# ECE 2195 Project Report

**Hardware AI acceleration with data pre-processing**

**Team member:**

Yue Tang [YUT51@pitt.edu](mailto:YUT51@pitt.edu)

Jinming Zhuang [JIZ230@pitt.edu](mailto:JIZ230@pitt.edu)

1. **Publicity**

We agree to include our project in the ECE2195 22Spring Class Project Website

1. **Problem Descriptions**

In recent years, various researchers have incorporated machine learning (ML) and convolutional neural network (CNN) to a wide range of applications in different domains including disease diagnosis [1][2][3], natural language processing (NLP) [4][5], electrocardiogram (ECG) signal classification [6][7], image dehazing [8] and so on. However, deep learning algorithms incur false-positive and false-negative cases when targeting the aforementioned applications. Pre-processing is deemed to be an effective method for improving the performance of DL models.

In this project, we plan to re-implement several state-of-art ML algorithms with and without pre-processing to verify the accuracy gain provided by the pre-processing steps. We have two tasks: GridDehazeNet for Image Dehazing and CNN based image classification with ZCA. For Task 1, we reproduce the GridDehazeNet algorithm training on the Indoor Training Set (ITS) of the RESIDE dataset [9] and testing on the Synthetic Objective Testing Set (SOTS) of the RESIDE dataset [9]. For Task2, we reproduce the experiment result of a 10-class image classification task based on cifar-10 dataset [13] with three kinds of preprocess methods mentioned in [10], including mean normalization, standardization and Zero Component Analysis(ZCA). Besides, in order to deploy the pre-processing and ML algorithms on a 16nm data-center AMD/Xilinx U200 board, we will build analytical models to guide the pre-processing and CNN kernel design.

1. **Background and Discussion on Prior work**

For different applications, several pre-processing methods are added to improve the inference performance. [1] and [2] used the Sobel filter and median filter before classifiers to improve the accuracy of disease detection. Before doing text classification, [4] took tokenization to break the raw text into words. [5] covers most of the common pre-processing techniques for NLP including lemmatization, stemming, handling negations, and so on. For ECG signal classification applications, [6][7] implement their filter to remove the noises under/above a certain frequency. For image dehazing, which can recover the clear version of a hazy image, [8] proposed an end-to-end trainable CNN network, named GridDehazeNet, which consisted of three modules: pre-processing, backbone, and post-processing. [10] proposed a ZCA image preprocessing technique which whitens the picture making the edges of the objects more prominent. They compared ZCA with mean normalization and standardization preprocessing techniques and find that ZCA outperformances both of them in terms of accuracy for image classification tasks on CNN models.

To implement CNN, [11] exploited various optimization techniques including loop unrolling, loop tiling, and loop transformation on the FPGA accelerator, and proposed a roofline model to quantitatively analyze its computing throughput and required memory bandwidth. [11] uses 6 parameters to present one layer of CNN which are N (# of input channel), M (# of output channel), R (# of row of output feature map), C (# of column of output feature map), K (length of kernel), S (# of stride). The maximum throughput of an application on a specific hardware platform can be presented by Eq. (1) in which Computational Roof and computation to communication (CTC) Ratio are defined by Eq. (2) and (3) respectively. BW refers to off-chip bandwidth. To increase the reuse rate of input feature map, weights and output feature map, on-chip buffers are allocated on FPGA. At the same time, tiling is applied to make the data fit on chip. In the following equation, Tn, Tm, Tr, and Tc refer to the tiling factor of N, M, R, C. αin, αout, αwght and Bin, Bout, Bwght refers to the number of access time and on-chip buffer size for input/output feature maps and weights respectively.

(1)

(2)

(3)

where

(4)

(5)

(6)

(7) (8)

In order to maximize the throughput, proper tiling factor should be applied under the DSP and on chip buffer constrains. To guide the design, following analytical model is proposed by [11]:

(9)

(10)

1. **Proposed Methodology**

In this project, we plan to implement several state-of-art ML algorithms and compare the inference performance with and without data pre-processing. The algorithms that are expected to be implemented are as follows.

Task 1. GridDehazeNet for Image Dehazing.

[8] proposed an end-to-end trainable CNN, named GridDehazeNet, for single image dehazing. The GridDehazeNet consists of three modules: pre-processing, backbone, and post-processing, which is illustrated in Figure 1. The pre-processing module contains a convolutional layer and a residual dense block (RDB) to effectively converts the single image dehazing problem into a multi-image dehazing problem by generating several variants of the given hazy image. The backbone is a combination of RDB blocks and upsampling/downsampling blocks. The post-processing module is also an RDB followed by a Conv layer, which aims to improve the quality of the dehazed image. The GridDehazeNet is trained on the Indoor Training Set (ITS) of the RESIDE dataset [9] and tested on the Synthetic Objective Testing Set (SOTS) of the RESIDE dataset [9]. The evaluation metric is peak signal to noise ratio (PSNR) and structure similarity (SSIM).



Figure 1 The architecture of GridDehazeNet[8]

Firstly, we reproduce the GridDehazeNet algorithm with and without the pre-processing module on GPU. The whole GridDehazeNet has 111 Convolution (Conv) layers. Followed by [8], to evaluate the performance of the network without pre-processing, we remove the pre-processing module and replace the first three learned inputs by the RGB channels of the given hazy image and the rest by all-zero feature maps. Such network has 105 Conv layers. The input image size of the testing dataset is 620460 ().

Then, we build analytical model to find the optimal design parameters to map the testing phase on U200 FPGA board. Followed by [11], the parameters we need to optimize include Tm, Tn for the whole network and Tr and Tc for each convolution layer. We use Eq. (11) to calculate the *computation clock cycles*. Then, we use the following equation to calculate the *communication clock cycles,* where *f* is the working frequency of the FPGA board, *DDR\_BW* is the theoretical DDR bandwidth, and the *total amount of external data access* is calculated by Eq. (11). In FPGA-based accelerator design, double buffer is adopted so that the computation and communication can be processed in parallel. Therefore, to estimate the total cycles of the whole Conv layer, we select the maximum of the *computation clock cycles* and the *communication clock cycles*.

(11)

It should be noted that, we have more than 100 layers to schedule the optimal parameters, the input image size is 620460, the maximum number of input channel is 128, and the maximum of output channel is 64. Therefore, complexity of searching the best Tm, Tn, Tr, and Tc is around 10012864620460≈2.341011, which costs so many times. Therefore, we need to make an optimization. We find that the value of the *computation clock cycles* is mainly dependent on the Tm and Tn, and the selection of Tr and Tc have little impact on it. As for the communication latency, it decreases when the TrTc increases. Under the same TrTc, the smaller number of Tc can decrease the continuity of data. Previous works have shown that the discontinuity of memory access can degrade the DMA transferring speed from about 8GB/s to around 1GB/s [12]. Therefore, we fix Tc=C to improve memory access continuity and reduce the searching complexity for GridDehazeNet.

Task 2. CNN based image classification with ZCA

In [10], ZCA has been proposed to preprocess images before sending input data to the CNN model while the mean normalization and standardization filter serves as the baseline for comparison. Mean normalization and standardization are defined by Eq. (12) and Eq. (13), where Xmean and Xstd are the corresponding normalized dataset, X is the oriental dataset, x\_m is the mean of X and ɜ is the standard deviation of X.

(12)

(13)

Eq. (14) defines ZCA preprocess, where XRGB is the dataset after RGB normalization described in Eq. (15). U and S are left singular vectors and singular values of the singular value decomposition (SVD) of XRGB.  e is a hyper-parameter deciding the whitening degree.

(14)

(15) We implement the CNN model proposed in [10] without any preprocessing technique. Then we applied several preprocess combinations on it including pure RGB normalization, RGB+mean normalization, RGB+standardization and ZCA. We also applied our analytical model on VGG16 network with ZCA preprocessing technique which is actually a matrix multiply layer. Compared with GridDehazeNet, VGG16 is relatively small, thus we can use exhaustive search to determine the number of Tn, Tm, Tr, Tc.

1. **Artifacts Description**

* The project code Github repo: <https://github.com/JinmingZhuang/ECE2195>.
* Hardware details: We first reproduce the GridDehazeNet algorithm of Task 1 with and without the pre-processing module and the CNN model of Task 2 with and without ZCA on the V-100 GPUs from the Pittsburgh Supercomputing Center (PSC) server to validate the PSNR and SSIM of Task 1 and classification accuracy of Task 2 with and without data preprocessing. Then, we build the analytical model to estimate the CNN inference latency under the hardware constraints of U200. The number of DSPs is 6833, the on-chip buffer size is 41.92MB, the datatype is floating point 32 bit, the clock frequency is 250MHZ, and the DDR bandwidth is 77GB/s.
* Software dependencies: For Task 1, we install Python 3.7 on the GPU with the CUDA version as 10.2. The GridDehazeNet is built on the Pytorch framework with the Pytorch version as 1.7.0 and Torchvision version as 0.8.0. Other necessary software packages are shown in our Github repo. For Task 2, we use Python 2,7 with opencv-python version as 4.2.0.32, theano version as 0.7. The python version analytical model can be successfully implemented with the same software dependencies. We also implement a Matlab version of analytical model which has the same result of the python version.
* Installation: The ITS training dataset and the SOTS testing dataset for Task 1 can be downloaded from RESIDE as described in GitHub repo. And Task 2 uses cifar-10 dataset which can be downloaded through our GitHub instruction. The required software packages can be installed by ‘conda’ or ‘pip’ command.
* Experiment workflow: In Task 1, we first train the GridDehazeNet with and without the pre-processing module. Same with [8], we train the model for 100 epochs, the learning rate is 0.001, the crop size is 240240. Due to the CUDA memory limitation, the batch size decreases from the original 18 to 16. Then, we test the well-trained model on the testing dataset. Same with [8], we use the original size of the image, which is 620460. After that, we build an analytical model to estimate the testing phase targeted on the U200 board. The GridDehazeNet with the pre-processing has 105 Conv layers. In Task 2, we train the model

with/without different combination of preprocessing techniques. We set learning rate to 0.05, lambda to 10.0 and train the model for 20 epochs. Then, we applied ZCA preprocessing technique before VGG16 network and use our analytical model to search for the design parameters. All the commands to run the code can be seen in the repo.

* Evaluation and expected results: For Task 1, after training for 100 epochs, the testing results of GridDehazeNet with the pre-processing module are PSNR as 32.12 and SSIM as 0.9833, while the results reported in the original paper are PSNR as 32.16 and SSIM as 0.9836. The results of GridDehazeNet without the pre-processing module are PSNR as 30.03 and SSIM as 0.9789, while the results reported in the original paper are PSNR as 31.48 and SSIM as 0.9820. From our analytical model, the optimal Tm and Tn for GridDehazeNet with and without pre-processing are both Tm=16 and Tn=80, which means the two situations can use the same FPGA design framework. As for the estimated latency, the GridDehazeNet with pre-processing takes up around 1.429\*108 clock cycles, while the network without pre-processing takes up around 1.297\*108 clock cycles, so the pre-processing module takes up around 9.23% of the total latency of GridDehazeNet. The results of every step are also uploaded in our repo. For Task 2, after 20 epochs the accuracy of baseline design without any preprocessing is 10.0%. The accuracy of RGB norm, RGB+mean, RGB+std, RGB+ZCA are 57.02%, 54.18%, 68.90%, 68.95% respectively. We use analytical model to measure the execution cycle with/without ZCA preprocessing. Without preprocessing, Tm=133, Tn=10, it takes 1.9136\*107 cycles. With ZCA preprocessing, Tm=136, Tn=10, it takes 2.965\*108 cycles for preprocessing and 1.9141\*107 cycles for CNN layers.

**Reference:**

[1]Beeravolu, Abhijith Reddy, et al. "Preprocessing of breast cancer images to create datasets for deep-cnn." IEEE Access 9 (2021): 33438-33463.

[2]Wang, Yu, and Shan Shan. "Accurate disease detection quantification of iris based retinal images using random implication image classifier technique." Microprocessors and Microsystems 80 (2021): 103350.

[3]Ferentinos, Konstantinos P. "Deep learning models for plant disease detection and diagnosis." Computers and electronics in agriculture 145 (2018): 311-318.

[4]Vijayakumar, T., and R. Vinothkanna. "Capsule network on font style classification." Journal of Artificial Intelligence 2.02 (2020): 64-76.

[5]Symeonidis, Symeon, Dimitrios Effrosynidis, and Avi Arampatzis. "A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis." Expert Systems with Applications 110 (2018): 298-310.

[6] Diker, Aykut, et al. "A novel ECG signal classification method using DEA-ELM." Medical hypotheses 136 (2020): 109515.

[7] Park, Yeonghyeon, Il Dong Yun, and Si-Hyuck Kang. "Preprocessing Method for Performance Enhancement in CNN-based STEMI Detection from 12-lead ECG." IEEE Access 7 (2019): 99964-99977.

[8] Liu, Xiaohong, et al. "Griddehazenet: Attention-based multi-scale network for image dehazing." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[9] Li, Boyi, et al. "Benchmarking single-image dehazing and beyond." IEEE Transactions on Image Processing 28.1 (2018): 492-505.

[10] Pal, Kuntal Kumar, and K. S. Sudeep. "Preprocessing for image classification by convolutional neural networks." 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). IEEE, 2016..

[11] Zhang, Chen, et al. "Optimizing fpga-based accelerator design for deep convolutional neural networks." Proceedings of the 2015 ACM/SIGDA international symposium on field-programmable gate arrays. 2015.

[12] Guan, Yijin, et al. "FP-DNN: An automated framework for mapping deep neural networks onto FPGAs with RTL-HLS hybrid templates." 2017 IEEE 25th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM). IEEE, 2017.

[13] Krizhevsky, Alex, and Geoffrey Hinton. "Learning multiple layers of features from tiny images." (2009): 7.