論文の要旨

題目 A Study on Machine Learning for Image Processing using GPUs
(GPU を用いた画像処理に対する機械学習の研究)

氏名　　松村 直樹

Machine learning is an application of artificial intelligence that makes systems learn the ability and improve from experience automatically without being explicitly programmed. In recent years, machine learning has been developing rapidly with the developing of *Convolutional Neural Networks (CNNs)*. To reduce the computation time of them, *Graphics Processing Units (GPU)* is mainly used for accelerator. In this dissertation, we research in machine learning for image processing using the GPUs.

In Chapter 1, we describe the introduction of the dissertation including the research background and contributions. We then go on to introduce the CNNs and its training flow in Chapter 2. The CNNs are composed of *convolutional layers* and *Fully-Connected Layers (FCLs)*, and applicable to various applications such as image classification and scene classification. Also, the CNNs are used for subroutines in image transformation networks such as *Generative Adversarial Networks* and *pix2pix*.

In Chapter 3, we describe NVIDIA GPU architecture and CUDA. Latest GPUs are designed for general purpose computing that include the computation of machine learning. They can perform computation in applications traditionally handled by the CPU.

In Chapter 4, we present two tile art image generation algorithms using greedy algorithm and machine learning as a greedy approach and a machine learning approach, respectively. The goodness of a generated tile art image is defined by the error between its projected image onto human eyes and the original image. The projected image is simulated on the computer using a two-dimensional Gaussian filter. From the goodness, the approach generates tile art images by pasting tiles one by one to the white canvas image. However, it takes a large amount of computation time, therefore, we implement the greedy approach on the GPU. The greedy approach on NVIDIA TITAN V GPU can run up to 318 times faster than that on CPU with single thread. The machine learning approach aims to generate tile art images by image

transformation networks. As the training dataset, we have used the tile art images generated by the greedy approach. Also, we have adopted the pix2pix as the tile art image generation network. As a training result, the network can generate a tile art image of size 4096×3072 within 1.04 seconds while the greedy approach on the GPU takes 571 seconds. However, in the tile art images generated by the machine learning approach, some tiles have lack of edge and noises that are not included in the greedy approach. Therefore, we propose an improvement technique of the machine learning approach to generate a high quality tile art image using iterative inference. As a result, in the tile art image with iterative inference technique, the characteristics of tiles can be enhanced.

In Chapter 5, we present a novel structured sparse FCL in the CNNs for image classification problem. The proposed approach eliminates the connections between the input and the output nodes except for those in the same position of the feature maps. Since the proposed architecture is defined initially, it is suitable to parallel computation. Therefore, we introduce an efficient implementation for the proposed sparse FCLs on the GPU. As a result for the large scale image recognition dataset, the proposed approach achieves a 14.7 times compression with 0.68% top-1 accuracy and 0.19% top-5 accuracy decrease for AlexNet network, and a 21.3 times compression with 0.68% top-1 accuracy and 0.31% top-5 accuracy decrease for VGG-16 network. Also, in the experiment on NVIDIA RTX 2080 Ti GPU, the GPU implementation for the proposed FCLs achieves speed-up factor 14.97 and 16.67 for forward and backward propagation compared to that for the non-compressed FCLs, respectively. Furthermore, to confirm that our proposed approach is applicable to practical image classification problems, we have trained the proposed models using transfer learning on various datasets. The experimental results show the proposed approach can achieve high test accuracy with high compression ratio on each dataset.

In Chapter 6, we conclude our works.