University of Victoria | Graduate Studies

Notice of the Final Oral Examination
for the Degree of Doctor of Philosophy

of

FABRIZIO PEDERSOLI

MSc (University of Brescia, 2012)
BSc (University of Brescia, 2009)

"Image Processing and Forward Propagation using Binary Representations and Robust Audio Analysis Using Deep Learning"

Department of Computer Science

Monday, March 11, 2019
1:30 P.M.
Clearihue Building
Room B007

Supervisory Committee:
Dr. George Tzanetakis, Department of Computer Science, University of Victoria (Supervisor)
Dr. Kwang Moo Yi, Department of Computer Science, UVic (Member)
Dr. Alexandra Branzan Albu, Department of Electrical and Computer Engineering, UVic (Outside Member)

External Examiner:
Dr. Leonid Sigal, Department of Computer Science, University of British Columbia

Chair of Oral Examination:
Dr. Kim Juniper, School of Earth and Ocean Sciences, UVic

Dr. David Capson, Dean, Faculty of Graduate Studies

Abstract

The work presented in this thesis consists of three main topics: document segmentation and classification into text and score, efficient computation with binary representations, and deep learning architectures for polyphonic music transcription and classification. Optical Character Recognition (OCR) and Optical Music Recognition (OMR) can be used to extract information from large collections of scanned documents. In the case of musical documents, an important problem is separating text from musical score by detecting the corresponding boundary boxes so that each process (OCR or OMR) can be applied to the correct type of data. Therefore, a new algorithm is proposed for pixel-wise classification of digital documents in musical score and text. It is based on a bag-of-visual-words approach and random forest classification. A robust technique for identifying bounding boxes of text and music score from the pixel-wise classification is also proposed.

For efficient processing of learned models, we turn our attention to binary representations. When dealing with binary data, the use of bit-packing and bit-wise computation can reduce computational time and memory requirements considerably. Efficiency is a key factor when processing large scale datasets and in industrial applications. For example OMR and OCR can benefit from efficient processing of binary images. SPmat is an optimized framework for binary image processing. We propose a bit-packed representation for binary images that encodes both pixels and square neighborhoods, and design SPmat, an optimized framework for binary image processing, around it. Using the SPmat representation, we define and evaluate optimized implementations of a variety of binary image processing algorithms such as: erosion/dilation, run-length extraction, contour extraction, and thinning. Bit-packing and bit-wise computation can also be used for efficient forward propagation in deep neural networks. Quantified deep neural networks have recently been proposed with the goal of improving computational time performance and memory requirements while maintaining as much as possible classification performance. In such networks, the weights and activations are quantized to lower precision and integer arithmetic is used to speed-up computations. A particular type of quantized neural networks are binary neural networks in which the weights and activations are constrained to -1 and $+1$. In this thesis, we describe and evaluate Espresso, a novel optimized framework for fast inference of binary neural networks that takes advantage of bit-packing and bit-wise computations. Espresso is self-contained, written in

C/CUDA and provides optimized implementations of all the building blocks needed to perform forward propagation. In the context of Espresso, we also describe how binary techniques can be used for efficient forward propagation of convolutional neural networks, a case not covered by existing literature on binary neural networks.

Following the recent success, we further investigate Deep neural networks. They have achieved state-of-the-art results and outperformed traditional machine learning methods in many applications such as: computer vision, speech recognition, and machine translation. However, in the case of music information retrieval (MIR) and audio analysis, shallow neural networks are commonly used. The effectiveness of deep and very deep architectures for MIR and audio tasks has not been explored in detail. It is also not clear what is the best input representation for a particular task. We therefore investigate deep neural networks for the following audio analysis tasks: polyphonic music transcription, musical genre classification, and urban sound classification. We analyze the performance of common classification network architectures using different input representations, paying specific attention to residual networks. We also evaluate the robustness of these models in case of degraded audio using different combinations of training/testing data. Through experimental evaluation we show that residual networks provide consistent performance improvements when analyzing degraded audio across different representations and tasks. Finally, we present a convolutional architecture based on U-Net that can improve polyphonic music transcription performance of different baseline transcription networks.