Abstract

The subject of this work is the domain of Convolutional Neural Networks. In this thesis, we analyze optimizations of Convolutional Neural Networks for low-resource devices. Given the increasing number of portable devices, the problem of creating lightweight machine learning models is of strong interest nowadays. The main challenges gravitate around the very limited environment these devices live in, such as low resources (battery, memory, CPU, GPU) and low network connectivity (low bandwidth is usually the case).

The thesis is structured in three chapters. Thr first chapter gives an introduction to artificial neural networks, starting from the supervised learning method all the way to convolutional neural networks, focusing on both the theoretical and practical concepts. The second chapter introduces new mobile-ready convolutional neural network architectures. The purpose of the last chapter is to demonstrate the applicability of convolutional neural networks for low resource devices and to study their performance in real life scenarios.

The main constribution of the thesis consists in introducing a new convolutional neural networks model, called SimpLeNet, trained using distillation for image tagging that can run on low-resource devices such as smartphones, smartwatches, tablets or TVs. Our major goal is that of preserving the performance of the convolutional neural network. For emphasizing the effectiveness of SimpLeNet, both in terms of model's size reduction, as well as in terms of classification accuracy, several experiments are performed on various data sets for image classification. Experiments performed on various data sets for image classification emphasize the effectiveness of SimpLeNet, both in terms of model's size reduction, as well as in terms of classification accuracy.

Our second contribution consists of designing and implementing an application capable of making real time image classification directly on the device, without accessing the internet. The application also tracks the time took in order to make inferences so we can exactly measure its performance.

The original part of this thesis is contained in Chapter 2 and was published in the research paper

Cosmin-Ionuţ Rusu and Gabriela Czibula, *Optimizing Convolutional Neural Networks for low-resource devices.* Proceedings of the 14th International conference on Intelligent Computer Communication and Processing, Cluj-Napoca, Romania, 2018, under review (**ISI Proceedings**).

This work is the result of my own activity. I have neither given nor received unauthorized assistance on this work.

Cluj-Napoca, 23.06.2018 **Rusu Cosmin-Ionuţ**