Master thesis : machine learning under resources constraints

Auteur : Greffe, Nathan
Promoteur(s) : Geurts, Pierre
Faculté : Faculté des Sciences appliquées
Diplôme : Master en ingénieur civil en informatique, à finalité spécialisée en "intelligent systems"
Année académique : 2018-2019
URI/URL : http://hdl.handle.net/2268.2/6798

Machine Learning under resource constraint: Abstract

Nathan Greffe

June 8, 2019

Nowadays, Machine learning on embedded devices, for example smartphones, is a popular topic. This arises from the growing concern of the public for data privacy and the general usefulness of running a service without the need of an external server.

Many methods exist to reduce the inference time of different algorithms but they are not often compared together or combined. The goal of this thesis is thus to offer a review of some of these methods. The scope of this work is limited to image classification using Convolutional Neural Networks on a Raspberry Pi 3B. CIFAR-10 was used as a dataset and out of the many benefits of embedded devices friendly CNNs, we limited ourselves to inference time. In other words, our goal was to classify images on CIFAR-10 as accurately as possible for a given inference time.

The methods investigated and our main conclusions are the following:

- We compared different architectures between each other and modified them to increase their performances. There, we managed to improve the error rates by adding Squeeze-and-Excitation blocks to existing MobileNetv1/v2 and MnasNet architectures.
- Based on recent works, we modified several pruning algorithms to adapt architectures by changing the number of channels per layer. This did not show promising results in our case. We suspect, however, that this is related to the dataset we used and might be worthwhile on bigger datasets like ImageNet.
- We used knowledge distillation on the architectures obtained from our search. Knowledge distillation takes profit of the predictions of a big network to help training a smaller one. This pushes the accuracies of some networks appreciably further.
- We tried to use Tensorflow's quantization to decrease the inference time of the previous architectures at moderate costs in accuracy. However, these methods are not mature as of now and did not give any result.

In conclusion, our initial objective of reviewing many methods and testing their interaction has been completed. Some of these methods showed rather shy results in our experiments on the CIFAR-10 dataset and the Raspberry PI 3B. We believe, however, that more significant improvements could be obtained in different settings, as shown in several publications. Our breadth-first study also allowed to highlight several directions that would deserve deeper exploration.