Machine Learning and Data Analysis journal paper template*

***F. S. Author*[1], *F. S. Co-Author*[2], *and F. S. Name*[1,2]**
author@site.ru; co-author@site.ru; co-author@site.ru
[1]Organization, address; [2]Organization, address

The paper investigates the problem of deep learning model selection. A method can consider different requirements to optimize performance on a wide range of hardware. The method allows to control trade-off between prediction quality and performance on the intermediate stage of model selection without additional model fitting. The method is based on a modification of a differential architecture search algorithm (DARTS) called DARTS-CC. The method can adapt to benefits and weaknesses of hardware, which is especially important for mobile devices with limited computation budget. To evaluate the performance of the proposed algorithm, we conduct experiments on the Fashion-MNIST and CIFAR-10 datasets using/simulating different hardware (e.g. ...) and compare the resulting architecture with architectures obtained by other neural architecture search methods.

DOI: 10.21469/22233792

1 Introduction

Discovering architectures for different problems is a difficult task. If you want to find an architecture that considers hardware properties the task becomes inhumanly hard. Optimizing for mobile devices is crucial because of their limited computation budget. With a wide variety of devices it is important to optimize searching not for one, but many different capabilities and requirements.

There are different approaches to finding architectures: evolutionary, RL (reinforced learning) and gradient-based. According to [7] gradient-based show similar prediction quality to other approaches but mainly with lower search cost.

On of the most popular algorithm is DARTS [5]. Main idea of the algorithm is to represent NN as DAG. Where each connection can be of the "candidate operations (e.g., convolution, max pooling, zero)"with different possibilities.

FBNet [3] uses this algorithm to optimize latency. Model shows great results decreasing FLOPS and CPU latency almost without decreasing accuracy.

DARTS-CC [2] uses this algorithm to optimize model complexity with trade-off control. This modification of DARTS can produce different architectures in one fitting run with control to trade-off between quality and complexity.

We propose a method that will combine approaches of DARST-CC [2] and FBNet [3] to obtain multiple architectures optimized for different requirements with different trade-offs between prediction quality and performance in one run.

2 Problem statement

$\approx$ same as DARTS-CC (details must discussed).

*The research was supported by the Russian Foundation for Basic Research (grants 00-00-0000 and 00-00-00001).