БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики и информатики

Кафедра дискретной математики и алгоритмики

Аннотация к магистерской диссертации

«Разработка алгоритмов обучения и исполнения нейронных сетей с использованием разреженного представления для задач оценки позы человека»

Мищенко Никита Валерьевич

Научный руководитель – кандидат технических наук
Белоцерковский А. М.

Научный консультант – научный сотрудник
Калиновский А. А.

Минск, 2020

РЕФЕРАТ

Магистерская диссертация, 46 с., 25 рис., 5 таб., 2 ф., 18 источников.

МАШИННОЕ ОБУЧЕНИЕ, КОМПЬЮТЕРНОЕ ЗРЕНИЕ, СВЕРТОЧНАЯ НЕЙРОННАЯ СЕТЬ, ОПТИМИЗАЦИЯ ВЕСОВ, РАЗРЕЖЕННОЕ ПРЕДСТАВЛЕНИЕ ВЕСОВ, ИСПОЛНЕНИЕ НА МОБИЛЬНЫХ УСТРОЙСТВАХ.

Объект исследования – проблема исполнения сложных сверточных нейронных сетей на мобильных устройствах. В частности, исследуются влияние использования разреженного представления весов в задачах машинного зрения.

Цель работы – изучить влияние разных способов прореживания весов сверточной нейронной сети на точность распознавания алгоритмов машинного зрения и на время работы сверточной сети на мобильном устройстве.

Методы проведения работы – изучение существующих методов оптимизаций обучения и исполнения сверточных нейронных сетей на мобильных устройствах. Разработка алгоритмов по минимизации количества весов сети разными стратегиями, а также разработка алгоритмов для быстрого исполнения слоев свертки. Проведение экспериментов на разработанном ПО, сравнительный анализ с обычным способом обучения и исполнения.

Результаты – новый способ оптимизации исполнения сверточных нейронных сетей на мобильных устройствам, позволяющий более чем в два раза улучшить производительность, почти не потеряв в точности решения задачи. Разработаны и реализованы на языке программирования Python алгоритмы по прореживанию весов сверточных слоев разными стратегиями. Разработаны и реализованы на языке программирования Metal и Swift алгоритмы по занулению одной стратегией весов сверточного слоя.

Область применения – различные задачи компьютерного зрения на мобильных устройствах, использующие нейронные сверточные сети.

ABSTRACT

Master thesis, 46 p., 25 fig., 5 tables, 2 formulas, 18 sources.

MACHINE LEARNING, COMPUTER VISION, CONVOLUTIONAL NEURAL NETWORK, WEIGHTS OPTIMIZATION, SPARSE WEIGHTS REPRESENTATION, MOBILE DEVICES INFERENCE.

Object of research – the problem of executing complex convolutional neural networks on mobile devices. In particular, the influence of sparse weights representation in machine vision tasks.

Objective – to study the influence of convolutional neural network different zeroing weights strategies on the accuracy of recognition in machine vision task. And to study the influence on the inference time of the convolutional network on a mobile device.

Methods – study of existing methods for optimizing training and inference of convolutional neural networks on mobile devices. Development of algorithms with different zeroing strategies, as well as development of algorithms for fast convolution layers inference. Conducting experiments on the developed software, comparative analysis with the usual method of training and inference.

Results – a new way to optimize convolutional neural network inference on mobile devices, which allows to double the performance, almost without any drops in the algorithm accuracy. Zeroing algorithms have been developed and implemented in the Python programming language with three different zeroing strategies. Optimized inference with zeroing was developed and tested in the Metal and Swift programming language.

Application area – various computer vision tasks, especially ones using convolutional neural network with inference on mobile devices.