УДК 519.688

КЛАССИФИКАЦИЯ ПРИЗНАКОВ НА ПОВЕРХНОСТЯХ ЦВЕТОВ С ПОМОЩЬЮ СЖАТЫХ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Сметанин А.А., Першуткин А.Э., Духанов А.В.

ФГАОУ ВО «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики», Санкт-Петербург, e-mail: artem_smetanin@niuitmo.ru, dukhanov@itmo.ru, pershutkin.alexei@gmail.com

Развитие технологий, а также пользовательский опыт в неумолимо расширяющейся сфере мобильных устройств стимулирует расширение областей применения наукоемких компьютерных технологий на мобильных устройствах, включая решение задач повышения урожайности растительных культур. Одной из компонент такого блока задач является своевременное распознавание их заболеваний на основе фотографий поверхностей листьев. Данная работа посвящена разработке и применению сжатых моделей машинного обучения для классификации признаков на разнородных массивах данных. В предложенном подходе применяется комбинация наиболее приспособленных к работе на мобильных устройствах методов машинного обучения для распознавания признаков на поверхностях объектов по фотографиям соответствующих поверхностей. Обучение производится для наиболее подходящей для таких устройств архитектуры MobileNetV2. В ходе исследования были проведены эксперименты, подтверждающие эффективность подхода на массиве данных растительных культур (цветов). Точность классификации на малых выборках составила 90%, а на массивах данных стандартных размеров достигла 96,3%. Также, благодаря квантизации обученных нейронных сетей, удалось достичь увеличения быстродействия и уменьшения размера мобильного приложения почти в два раза по сравнению с неквантизированной моделью с допустимыми потерями качества классификации.

Ключевые слова: распознавание, классификация, сверточные нейронные сети, обучение с переносом, сиамские нейронные сети

CLASSIFICATION OF FEATURES ON FLOWER SURFACES USING COMPRESSED MACHINE LEARNING MODELS

Smetanin A.A., Pershutkin A.E., Dukhanov A.V.

Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, Saint Petersburg, e-mail: artem_smetanin@niuitmo.ru, dukhanov@itmo.ru, pershutkin.alexei@gmail.com

The development of technologies, as well as user experience in the inexorably expanding field of mobile devices, stimulates the expansion of applications of high-tech computer technologies on mobile devices, including solving problems of increasing crop yields. One of the components of such a block of tasks is the timely recognition of their diseases based on photographs of leaf surfaces. This work is devoted to the development and application of compressed machine learning models for the classification of features on heterogeneous data arrays. The proposed approach uses a combination of the most adapted machine learning methods to work on mobile devices to recognize features on the surfaces of objects from photographs of the corresponding surfaces. Training is performed for the MobileNetV2 architecture that is most suitable for such devices. In the course of the study, experiments were carried out confirming the effectiveness of the approach on a data array of plant crops (flowers). Classification accuracy on small samples was 90%, and on data arrays of standard sizes reached 96.3%. Also, thanks to the quantization of trained neural networks, it was possible to achieve an increase in performance and a reduction in the size of the mobile application by about two times compared to an unquantized model with acceptable loss of classification quality.

Keywords: recognition, classification, convolutional neural networks, transfer learning, siamese neural networks

В настоящее время распознавание признаков на разнородных поверхностях с использованием мобильных устройств становится все более актуальным. Развитие технологий, а также пользовательский опыт в расширяющейся сфере мобильных устройств стимулирует развитие все более новых технологий, рассчитанных на мобильные устройства в самых разных областях народного хозяйства, в том числе в задачах сбережения различных видов цветов и растений. Однако на данный момент технологии по обработке изображений на базе мобильных устройств зачастую не приспособлены к мобильным устройствам в силу ограничений по вычислительным ресурсам последних. Также немалую роль играет возможность сбора достаточного объема набора данных для создания качественных моделей машинного обучения. На сегодняшний день диагностика болезней растений, идентификация повреждений на поверхностях нуждаются в участии человека, что вызывает соответствующие издержки на вызовы специалистов в район работ.

Современный пользовательский опыт с каждым годом все больше обращен в сторону мобильных технологий. Это, в свою очередь, порождает бурный рост количества исследований в области адаптации

современных технологий к использованию на мобильных устройствах, включая как облачные сервисы, так и самостоятельные решения, работающие непосредственно на мобильном устройстве. В совокупности с трендом к цифровизации всех отраслей народного хозяйства, пользовательский опыт и рост числа исследований позволяют говорить о необходимости и возможности разработки технологий классификаций признаков на разнородных поверхностях.

Еще одной важной задачей является разработка такой технологии, которая могла бы эффективно функционировать на самых разнородных, но сходных морфологически датасетах. Таким образом, возникает задача разработки технологии машинного обучения, позволяющей осуществлять эффективную классификацию признаков на поверхностях с использованием малых разнородных выборок для обучения.

Целью исследования является разработка подхода к формированию сжатых моделей машинного обучения для высокоточной классификации объектов по их оптическим изображениям на основе малых выборок. Получаемые модели должны быть пригодны в том числе для мобильных устройств. Тем самым будет возможно решать задачи распознавания объектов и/или их свойств с помощью смартфонов, планшетов и иных видов мобильных устройств.

Материалы и методы исследования

Ранее такие задачи интеллектуальной обработки, как классификация изображений, решались в том числе и классическими методами машинного обучения [1, 2]. Позднее, с прорывом в области изучения когнитивных и обобщающих способностей нейронных сетей, на арену методов обработки изображений вышли сверточные нейронные сети, заняв на конкурсе ImageNet в 2012 г. первое место [3]. С тех пор сверточные нейронные сети постоянно совершенствовались, демонстрируя все более высокие показатели точности классификаций изображений. Появились такие архитектуры нейронных сетей, как ResNet [4], U-net [5], MobileNetV2 [6] и др. Обучение этих моделей машинного обучения составляло нередко довольно трудоемкий и ресурсоемкий процесс: один только датасет ImageNet состоял в 2019 г. из 14 миллионов изображений, принадлежащих различным классам объектов, и со временем датасет становится все больше [7], а время обучения нередко сопровождается значительными временными затратами, даже с использованием технологий параллельных вычислений на видеокартах. Для ряда задач уже существует способ избежать вычислительных издержек путем обучения нейросетей методом transfer learning [8]. Однако в силу небольшого объема либо полного отсутствия в ImageNet изображений, соответствующих области решаемой задачи (например, задаче идентификаций болезней растений, цветов), так или иначе возникает необходимость дообучения готовой модели. Это не всегда приводит к ожидаемому результату [9], точность классификации не превысила 87% при отсутствии возможности эффективного использования обученной модели на мобильном устройстве.

Предыдущие исследования

Чтобы избежать переобучения, а также необходимости создавать очень большой по величине датасет, ранее мы применили комплексный метод, основанный на обучении готовой модели transfer-learning методом с применением triplet-loss ошибки [9].

Ранее разработанная технология представлена схемой на рис. 1, на которой отражены следующие процессы:

– обучение сиамской нейронной сети-экстрактора признаков с применением triplet-loss-функцией ошибки;

– сжатие обученной модели с помощью применения методов квантизации;

– обучение классификатора признаков многослойного перцептрона.

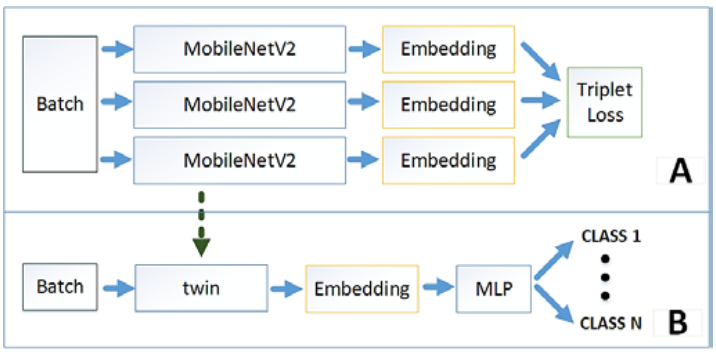

Рис. 1. Схема технологии создания сжатой модели – классификатора признаков

Реализованная нами ранее модель тестировалась на специально собранном датасете растений PDD [9]. Результаты эксперимента убедительно доказали применимость технологии для классификации признаков на поверхностях растений. Точность модели с применением таких подходов, как transfer-learning, сиамских нейронных сетей, а также обучение с применением функции ошибки triplet-loss, составила 99,5%. Применяемая архитектура: MobileNetV2.

Ключевой особенностью технологии является возможность one-shot обучения на выборках малого объема и применение квантизации модели с целью уменьшения ресурсоемкости процесса обучения.

Квантизация

Использовались следующие виды квантизации: статическая и динамическая.

Статическая квантизация заключается в замене весов модели, хранящихся, как правило, в формате float, занимающем в памяти 4 байта, на значения типа int8, занимающем в памяти лишь 1 байт (табл. 1) в соответствии с формулой (*).

Таблица 1

Ресурсозатраты на значения в разных типах данных

	Память	Такты процессора
INT8	2 байта	1–3 такта
FLOAT32	8 байт	2–8 тактов

Конкретные пределы минимальных и максимальных значений конкретного веса определяются с помощью модулей-наблюдателей, накапливающих статистику в ходе обучения о минимальных и максимальных пределах, в которых находилось значение веса за весь период обучения.

$$Q(x, scale, zero\_point) =$$

$$= round\left(\frac{x}{scale} + zero\_point\right), \quad (*)$$

где x – исходное значение веса,

scale – масштабирующий коэффициент,

zero_point – нулевой сдвиг.

Общий алгоритм статической квантизации модели выглядит следующим образом:

1) в исходной модели выбираются области с весами, за которыми закрепляются модули-наблюдатели;

2) производится обучение модели, в процессе которого наблюдатели определяют диапазоны весов;

3) по формуле (*) вычисляются целочисленные значения весов.

Сигнал в сети распространяется следующим образом:

1) на вход модели подаются квантизированные значения в формате int8;

2) выходные значения (inferences, в случае квантизации только сверточной подсети – экстрактора признаков) модели преобразуются вновь в значения float.

Динамическая квантизация включает в себя статическое (преобразование весов в int8), а также квантизацию активаций модели «на лету», что и объясняет динамичность процесса. Поскольку в иных видах квантизации преобразуются к int8 только веса модели, а сигналы, образующиеся на выходах активаций, остаются вещественными float-переменными, то и передаваемые между слоями сигналы все же остаются вещественными числами. Квантизация активаций позволяет получать на их выходах уже значения int8, что должно способствовать ещё большему ускорению и сжатию модели.

Эксперимент

С целью оценки точности классификации на датасете, отличном от ранее использованного датасета PDD, был проведен эксперимент по классификации изображений на датасете с изображениями поверхностей цветов.

Основной задачей эксперимента является также и оценка применимости модели на мобильных устройствах, а также меньшие ресурсозатраты при формировании обучающих датасетов и обучения модели в целом. Ключевыми факторами, напрямую влияющими на ресурсозатратность всего процесса обучения, являются:

1. Величина обучающей выборки.
2. Размер готовой сжатой модели в Мб.
3. Скорость работы сжатой модели.

Для оценки потерь в точности классификации при обучении малыми выборками, мы провели сравнительные эксперименты не только на полном наборе обучающих данных, но и на сокращенном наборе данных: Flower Recognition, Flower Recognition, Small Flower Recognition соответственно.

Flower Recognition – это открытый набор данных с изображениями цветов, загруженных на платформу Kaggle. В этом наборе данных 4242 изображения цветов (пример на рис. 2). Сбор данных основан на данных flickr, google images, yandex images. Набор данных содержит такие классы, как ромашка, тюльпан, роза, подсолнух, одуванчик. Для каждого класса есть около 800 фотографий. Фотографии имеют невысокое разрешение, около 320x240 пикселей. Для экспериментов были определены два

типа этого набора данных. Для одного эксперимента был взят полный набор данных, для другого набор данных был обрезан, и для каждого из классов было сделано менее 50 изображений.

Основные параметры датасетов, в том числе и параметры ранее применявшегося датасета PDD, приведены в табл. 2.

Была произведена серия экспериментов для каждого из наборов данных с применением и без применения методов квантизации. Результаты экспериментов приведены в табл. 3.

Результаты исследования и их обсуждение

В ходе экспериментальных исследований выяснилось, что точность классификации на новом, несжатом датасете незначительно ниже показателя точности для модели, обученной на первоначальном обучающем наборе. Чуть больший отрыв по качеству классификации у PDD и Small Flower Dataset – сокращенной версии исходного датасета: 87–90%. Однако важно учесть, что:

1. PDD более чем в два раза превосходит Small Flower Dataset по количеству изображений.

2. Наименьшие показатели точности, полученные для квантизированных моделей, являются достаточными для эффективного применения обученных моделей на мобильных устройствах в полевых условиях.

Среди методов квантизации наиболее эффективным оказался метод статической квантизации: при практически полном отсутствии разницы в показателях точности классификации с моделью, полученной с применением динамической квантизации, размер модели, полученной с применением статической квантизации, отличается почти в два раза, а время, затрачиваемое на обработку ста изображений, меньше почти в семь раз. Это объясняется, прежде всего, наличием дополнительных блоков, предназначенных для реализации динамической квантизации активации моделей.

Рис. 2. Пример датасета с изображениями поверхностей цветов

Таблица 2

Параметры массивов данных

Наименование	Количество изображений	Количество классов	Краткое описание
PDD	611	15	Фотографии листьев больных и здоровых растений
Flower Recongnition	4242	5	Фото цветов из интернета
Small flower recongition	250	5	Сокращенная версия датасета с изображениями цветов

Таблица 3

Показатели точности классификации и ресурсоемкости процесса обучения квантизованных моделей в сравнении с исходными

Метрика	Без квантизации	Динамическая квантизация	Статическая квантизация
Точность на датасете PDD	98%	98%	98%
Точность на Flower dataset	96,3%	94,2%	96,4%
Точность на значительно уменьшенном Flower dataset	90%	88%	87%
Время, затраченное на обработку 100 изображений, с	14,2	14,2	2,6
Размер модели, Мб	18,6	13,2	7,6

Таким образом, технология обучения сжатых моделей применима не только к задаче классификации признаков на растениях с целью идентификации признаков на поверхностях растений, но также и для классификации признаков на поверхностях цветов, что и подтверждают показатели точности, приведенные в табл. 3.

Практическая применимость

В ходе разработки технологии был разработан Телеграм-бот и мобильное приложение Android.

Телеграм-бот представляет собой интерфейс, посредством которого пользователь может получить топ-5 результатов прогнозирования, по убыванию вероятности принадлежности изображения к тому или иному классу (рис. 3).

Модель, используемая в боте, запускается в Docker-контейнере. Приложение реализовано с помощью Python с использованием Telegram API tools.

Мобильное приложение для Android представляет собой уже самостоятельную программную единицу, результаты вычисления которой уже не зависят от наличия соединения с сервером, осуществляющим расчеты, предоставляющего интерфейс и т.д. Загрузив данное приложение на мобильное устройство, пользователь получает возможность офлайн-классификации загруженного изображения растения или цветка из галереи (рис. 4).

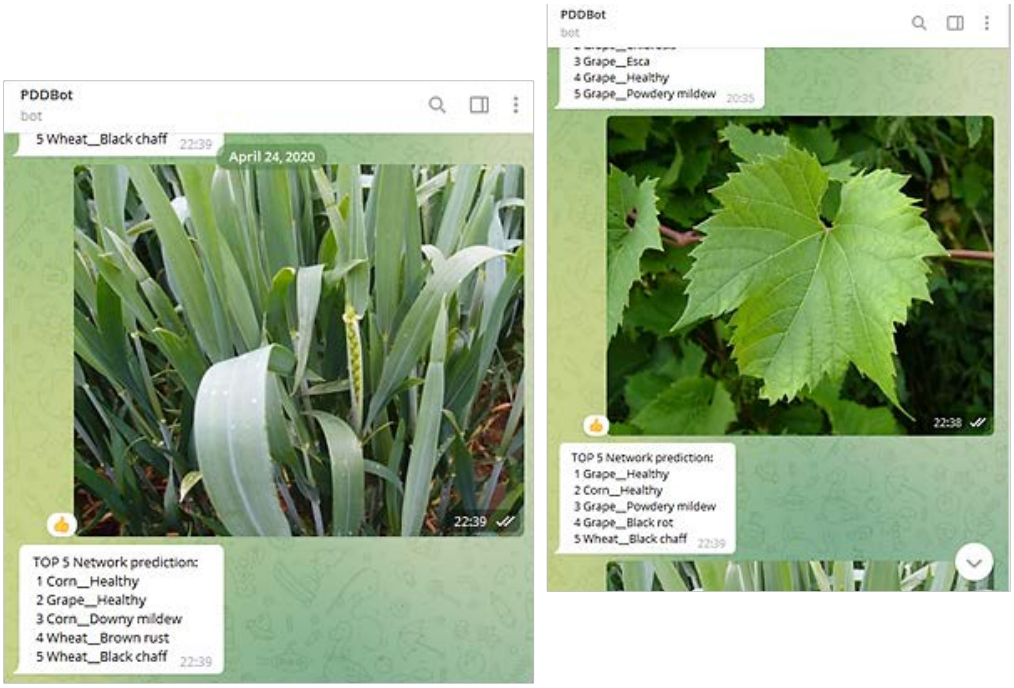

Рис. 3. Скриншоты, демонстрирующие работу с Телеграм-ботом

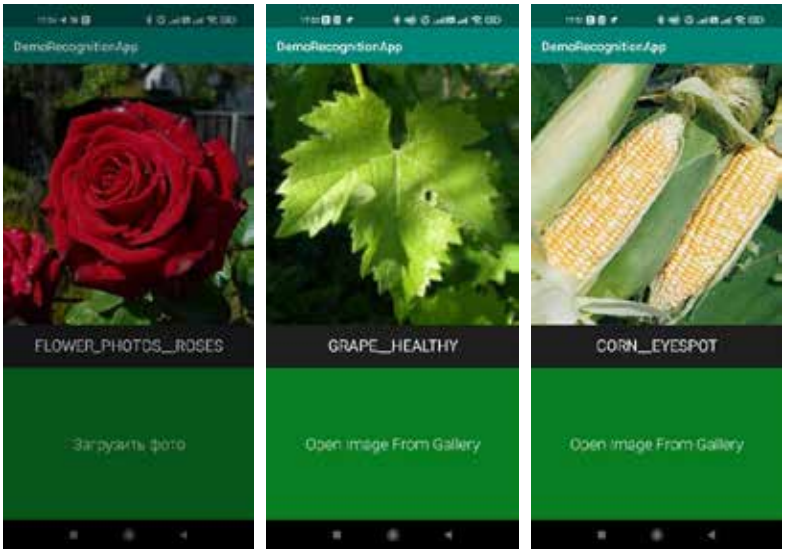

Рис. 4. Скриншоты, демонстрирующие работу приложения

Реализация мобильного приложения осуществлялась на языке программирования Java в Android Studio. Рабочий экземпляр приложения на устройстве включает в себя статически квантизированную модель сверточной триплет-сети-экстрактора признаков и классификатор на основе многослойного перцептрона.

Имея на своем устройстве мобильное приложение, пользователь получает возможность офлайн классификации изображений малой (квантизированной) моделью. В то же время у пользователя остается возможность более точной классификации изображений неквантизированными моделями посредством Телеграм-бота при наличии сети Интернет.

Заключение

В данной работе предложен и протестирован подход к формированию сжатых моделей машинного обучения для высокоточной классификации объектов по их оптическим изображениям на основе малых выборок. Здесь применена модель машинного обучения, созданная с применением комбинаций таких методов машинного обучения, как transfer-learning, Сиамская сеть, с обучением классификатора с MLP и трехчленной функцией ошибки. Эти методы показали свою эффективность как для небольших наборов данных (90%), так и для наборов с классами, сходными по морфологии с ранее использовавшимся датасетом PDD (96,3%).

В будущем планируется продолжить исследования, рассмотреть такие архитектуры, как, например, U-net, RNN, а также провести больше экспериментов с наборами данных, схожих по морфологии.

Исследование выполнено в рамках НИР Университета ИТМО «Разработка технологии формирования моделей машинного обучения малого объема в целях распознавания признаков на изображениях с высокой точностью при сверхмалых выборках данных».

Список литературы

1. Sheykhmousa M., Mahdianpari M., Ghanbari H., Mohammadimanesh F., Ghamisi P., Homayouni S. Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2020. Vol. 13. P. 6308–6325.

2. Друки А.А., Спицын В.Г., Болотова Ю.А., Башлыков А.А. Семантическая сегментация данных дистанционного зондирования Земли при помощи нейросетевых алгоритмов // Известия Томского политехнического университета. Инжиниринг георесурсов. 2018. Т. 329. № 1. С. 59–68.

3. Krizhevskiy A., Sutskever I., Hinton G.E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012. Vol. 25.

4. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. P. 770–778.

5. Ronneberger O., Fischer P., Brox T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer. Cham, 2015. P. 234–241.

6. Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. C. 4510–4520.

7. Iorga C., Neagoe V.E. A deep CNN approach with transfer learning for image recognition. 2019 11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). IEEE. 2019. P. 1–6.

8. Zhuang F., Qi Z., Duan K., Xi D., Zhu Y., Zhu H., He Q. A comprehensive survey on transfer learning. Proceedings of the IEEE. 2020. Т. 109. № 1. С. 43–76.

9. Сметанин А.А., Гончаров П.В., Ососков Г.А. Выбор методов глубокого обучения для решения задачи распознавания болезней растений в условиях малой обучающей выборки // Системный анализ в науке и образовании. 2020. № 1. С. 30–38.