УДК 004.932 DOI: 10.17587/it.26.586-593

А. В. Савченко, д-р техн. наук, проф., e-mail: avsavchenko@hse.ru,
И. С. Гречихин, аспирант, ст. преподаватель, e-mail: igrechikhin@hse.ru,
Национальный исследовательский университет Высшая школа экономики, Нижний Новгород

Детектирование специализированных категорий объектов на фотографиях в мобильных устройствах на основе многозадачной нейросетевой модели[1]

Предложен метод детектирования категорий нескольких различных видов объектов на фотографиях в мобильных устройствах. Вначале с использованием известных нейросетевых детекторов выделяются искомые объекты. Их характерные признаки извлекаются с помощью многозадачной нейросетевой модели с несколькими выходными слоями — по одному на каждый вид объекта. Представлены экспериментальные результаты для распознавания пород собак и кошек и группировки фотографий одного и того же животного.

Ключевые слова: *обработка изображений, сверточные нейронные сети, мобильные системы, распознавание пород животных, иерархическая кластеризация*

Введение

Задача определения предпочтений пользователей мобильных устройств в настоящее время становится все более актуальной в связи с непрерывным развитием рекомендательных систем. Один из вариантов ее решения связан с обработкой фотографий, сделанных самим пользователем. Как отмечено в работе [1], наибольший интерес представляют алгоритмы детектирования объектов на изображениях (предметов интерьера, видов еды, транспорта, спортивных принадлежностей, музыкальных инструментов и т.п.). Нередко требуется получить более специализированную информацию о предпочтении, в частности, определить подкатегории некоторых значимых классов (марки автомобилей, породы животных). К сожалению, современные наборы данных, предназначенные для обучения нейросетевых детекторов, не содержат данных о подкатегориях, либо существующих в этих наборах подкатегорий недостаточно для надежного обнаружения соответствующих им объектов.

Стоит отметить, что для многих важных подкатегорий доступны специализированные наборы данных, которые можно использовать для обучения классификаторов, например, глубоких сверточных нейронных сетей (СНС) [2]. Поэтому в настоящей работе используется двухэтапная процедура обнаружения объектов, в которой вначале для обнаружения более общих видов объектов применяются традиционные нейросетевые детекторы [3], а далее для нахождения специализированных подкатегорий применяется СНС. При этом особенность предлагаемого метода состоит в применении единой многозадачной СНС с несколькими выходами (по одному — для каждого вида объекта) [4]. В результате применения такого подхода можно не только снизить вычислительную сложность за счет отказа от применения нескольких СНС, но и использовать выходы промежуточных слоев сети в качестве характерных признаков анализируемых объектов, например, для группировки различных фото-

[1] Статья подготовлена в ходе проведения исследования (№ 19-04-004) в рамках Программы "Научный фонд Национального исследовательского университета "Высшая школа экономики" (НИУ ВШЭ)" в 2019—2020 гг. и в рамках государственной поддержки ведущих университетов Российской Федерации "5-100".

графий одного и того же объекта. В качестве примера реализации такого подхода в работе рассматривается классификация и последующая кластеризация различных видов домашних животных (пород кошек и собак) [5, 6]. Полученные результаты и сделанные по ним выводы рассчитаны на широкий круг специалистов в области распознавания образов.

1. Постановка задачи

Задача анализа предпочтений по фотографиям состоит в том, чтобы по поступившему на вход фотоальбому выделить наиболее интересные для пользователя категории из заранее заданного списка [1]. Результатом анализа предпочтений можно считать частоты встречаемости объектов каждой категории в фотоальбоме. Если для каждой категории задано множество изображений, соответствующих данной категории объектов, а также данные об их местонахождении на изображении (обрамляющие прямоугольники или маска границ), можно решить задачу с помощью обучения одного из современных высокоточных нейросетевых детекторов [7]. Архитектуры Faster R-CNN [3] также используют СНС для создания карты признаков, но с их помощью определяются несколько (100...200) регионов, в которых могут содержаться потенциально интересные объекты. После этого на основании карты признаков и выделенных регионов предсказывается класс объекта. В совокупности такая архитектура обнаруживает объекты значительно точнее за счет снижения вычислительной эффективности. Детектор SSD (Single Shot Detector) использует карту признаков на выходе СНС для предсказания классов и положения объектов за один проход, а его модификация SSDLite [8] включает разделяемые по глубине (depth-separable) сверточные слои для снижения вычислительной сложности и затрат памяти, что делает их удобными для использования в мобильных устройствах. Среди моделей, осуществляющих детектирование за один проход, следует выделить RetinaNet [9], которая позволяет за счет специальной функции потерь (focal loss) достичь достаточно высоких показателей точности и вычислительной эффективности.

Прорывом в области создания СНС, приемлемых как для классификации, так и для извлечения карт признаков, подходящих для использования в детекторах SSD и Faster R-CNN, стали архитектуры ResNet и Inception [10], которые сумели достичь высокого качества классификации изображений на значительном по размеру наборе изображений и категорий объектов. Для использования в мобильных устройствах [1], где есть ограничения по объему памяти и процессорной мощности, необходимы более вычислительно эффективные архитектуры, такие как MobileNet [8].

2. Многозадачные нейронные сети для классификации подкатегорий

К сожалению, во многих случаях сбор необходимого для обучения детектора набора данных оказывается слишком сложным. В частности, основная трудность состоит в получении разметки, необходимой для обучения детектора. Для этого на каждом изображении из обучающей выборки требуется указать область искомого объекта, чаще всего, с помощью выделения обрамляющего прямоугольника. При этом для получения высокой точности требуются сотни размеченных примеров каждого класса, и чем больше различных категорий, тем больше должно быть примеров объектов каждой категории.

В таком случае можно воспользоваться двухэтапной процедурой, в которой вначале с помощью нейросетевого *детектора* находятся $N > 1$ более общих видов объектов (автомобиль, еда, музыкальный инструмент, домашнее животное), а потом для каждого n-го вида ($n = 1, 2, ..., N$) выделяются специализированные категории. Пусть для n-го вида имеются $C_n > 1$ категорий. По результатам детектирования находятся обрамляющие прямоугольники для объекта n-го вида, после чего для каждого такого прямоугольника из изображения вырезается часть, принадлежащая найденному объекту. Далее выделенный объект распознается с помощью отдельного (для каждого n) *классификатора*, например, СНС.

Конечно, в этом случае время принятия решений может увеличиться за счет появления второй СНС. Однако такие классификаторы можно обучить, используя набор фотографий каждой подкатегории объектов n-го вида, в котором не требуется указывать обрамляющие прямоугольники, что существенно упрощает процедуру сбора и разметки данных. Более того, постоянно развиваются методы дообучения СНС на сверхмалых обучающих выборках (даже с одним примером каждой категории) [11], в то время как обучение части сети-детектора, следующей после извлечения карты

признаков, все еще требует больших объемов обучающих данных.

В связи с тем, что в процессе принятия решений исходные фотографии подаются на вход нейросетевого детектора, для обучения классификаторов также должен использоваться не исходный набор, а части изображений, полученные с помощью аналогичной процедуры выделения обрамляющих прямоугольников на выходе обученного детектора. Рассмотрим подробнее различные способы построения классификаторов на основе СНС.

Наиболее простой вариант — обучить отдельные классификаторы для каждого вида объекта. В настоящий момент в рамках технологии переноса знаний (transfer learning) [2] наиболее часто для настройки классификатора применяется не доступное обучающее множество, а сверхбольшая коллекция дополнительно собранных изображений, например, ImageNet. Такая коллекция используется для обучения глубокой СНС, состоящей из нескольких чередующихся слоев свертки и подвыборки, выход которых поступает на вход последовательно соединенных полносвязных слоев. Выход последнего сверточного слоя является четырехмерным тензором, поэтому далее обычно добавляется слой глобального усреднения (global average pooling) по ширине и высоте, после чего его выход из $D \gg 1$ значений поступает на вход последнего полносвязного слоя, в котором и принимается решение в пользу одной из подкатегорий. Такую архитектуру можно рассматривать как применение логистической регрессии (последний слой СНС) для классификации вектора $\mathbf{x}$ из D характерных признаков, выделенных на предыдущих слоях. Поэтому обычно последний полносвязный слой заменяется на новый слой с C_n выходами $\mathbf{z}_c$ (по одному на каждую подкатегорию n-го вида), в котором с помощью слоя softmax оцениваются апостериорные вероятности p_c принадлежности входного объекта c-й подкатегории ($c = 1, 2, ..., C_n$). После этого происходит дообучение (fine-tuning) полученного таким образом нейросетевого классификатора для доступного n-го обучающегося множества [2].

Такой способ является наиболее приемлемым, если для каждого вида доступно репрезентативное обучающее множество, при этом сами типы объектов существенно отличаются друг от друга. К сожалению, затраты памяти линейно зависят от числа видов N, при этом точность обученного классификатора может оказаться достаточно низкой, если имеется обучающая выборка малого размера. При этом, если необходимо добавить новый $(N + 1)$-й вид объектов, придется заново обучать новый классификатор.

Для преодоления указанных недостатков может применяться единая СНС, обученная для одновременного решения нескольких задач. Например, можно объединить все подкатегории в одно обучающее множество, состоящее из $(C_1 + C_2 + ... + C_N)$ классов. При этом необходимо выполнить дополнительную постобработку результатов классификации: использовать оценки апостериорных вероятностей, соответствующих только виду объекта, найденного детектором. Такой подход позволяет поддерживать всего один классификатор (рис. 1), однако при его обучении изображения одного вида оказывают влияние на выходы, ответственные за подкатегории других видов, что может привести к снижению итоговой точности.

Поэтому наиболее приемлемым способом реализации многозадачной СНС является использование одной сети для извлечения вектора признаков $\mathbf{x}$, который подается на N выходных слоев (heads) — по одному для каждого вида объекта. Такой подход (рис. 2) позволяет создать разные классификаторы на базе одной общей архитектуры СНС, которая получает карты признаков их входных изображений и передает одному из выходов для классификации. К недостаткам такого подхода можно отнести усложнение процесса обучения: веса нейронной сети, передающие информацию от карт признаков к каждому из выходов, для более сбалансированного обучения модифицируются отдельно в рамках итеративной процедуры для подвыборок объединенного обучающего множества (mini-batch), каждая из которых включает только один вид объектов [4].

Отметим, что на практике описанные выше подходы могут комбинироваться в гибридные архитектуры, если есть несколько схожих ви-

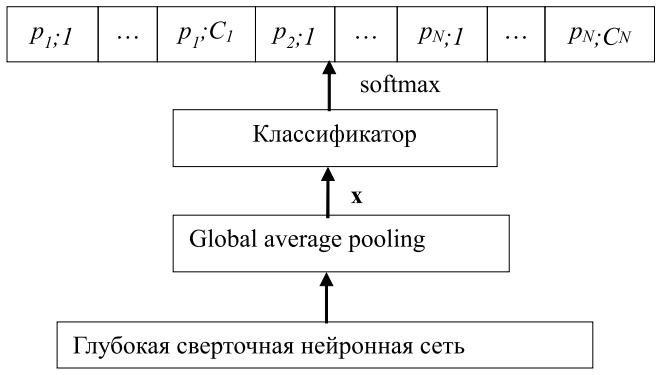

Рис. 1. Сверточная нейронная сеть с объединением всех подкатегорий

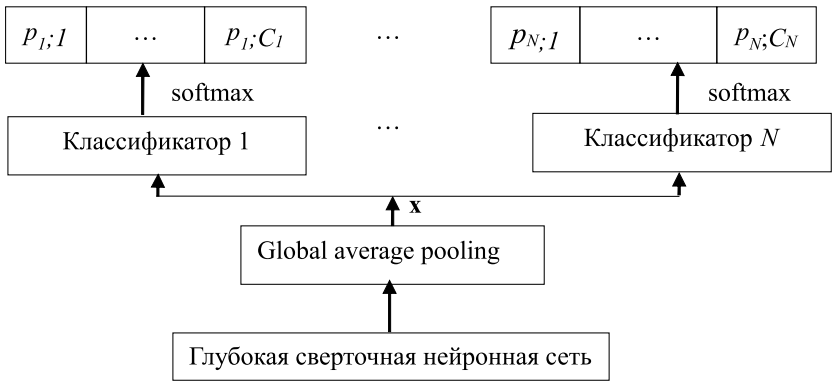

Рис. 2. Многозадачная сверточная нейронная сеть с N выходными слоями

дов объектов (например, породы котов и собак), которые существенно отличаются от других видов (например, автомобилей или видов еды). В таком случае использование единой базовой СНС для извлечения характерных признаков может оказаться неприемлемым, поэтому нужно использовать несколько независимых архитектур вида (рис. 2).

3. Предложенный подход

На рис. 3 представлена функциональная схема предлагаемой информационной системы извлечения предпочтений на основе детектирования категорий нескольких различных видов объектов. Здесь для каждой фотографии на первом этапе осуществляется детектирование общих видов объектов, специализированные категории которых на втором этапе предсказываются с помощью многозадачной СНС (см. рис. 2).

Кроме того, такая СНС применяется также для группировки фотографий идентичных объектов. Действительно, в галерее пользователя могут встречаться несколько фотографий одного и того же животного. Можно сделать предположение, что это животное представляет интерес для пользователя, например, может являться его домашним животным.

Для группировки объектов обычно применяются методы кластеризации [12]. Для того чтобы использовать алгоритмы кластеризации, каждый объект должен быть представлен некоторым числовым вектором. В частности, могут быть использованы вектор признаков **x** или вектор **p** апостериорных вероятностей классов p_c на выходе СНС. Так как число кластеров заранее не известно, в настоящей работе использовались методы агломеративной иерархической кластеризации и плотностной кластеризации (DBSCAN, HDBSCAN) [13], не требующие наличия информации о числе групп K. Первая группа методов создает иерархическую структуру кластеров, что может оказаться полезным при определении принадлежности кластеров. Для ее применения необходимо подобрать несколько гиперпараметров: мера близости, межкластерное расстояние и порог определения кластеров. Наилучшие параметры DBSCAN зависят от выборки и могут меняться в зависимости от набора векторов, поскольку этот алгоритм определяет кластеры одинаковой плотности. Метод HDBSCAN преодолевает это ограничение, однако он более вычислительно требователен [13].

Описанный алгоритм был реализован авторами в Android-приложении [14] анализа пользовательских предпочтений для мобильных устройств [1]. В приложении показываются фотографии из галереи пользователя, детектируются несколько видов объектов (автомобили, кошки и собаки), классифицируются марки автомобилей и породы домашних животных (рис. 4, см. третью сторону обложки). На странице со статистикой найденных категорий отмечаются как породы животных, так и статистика по найденным кластерам.

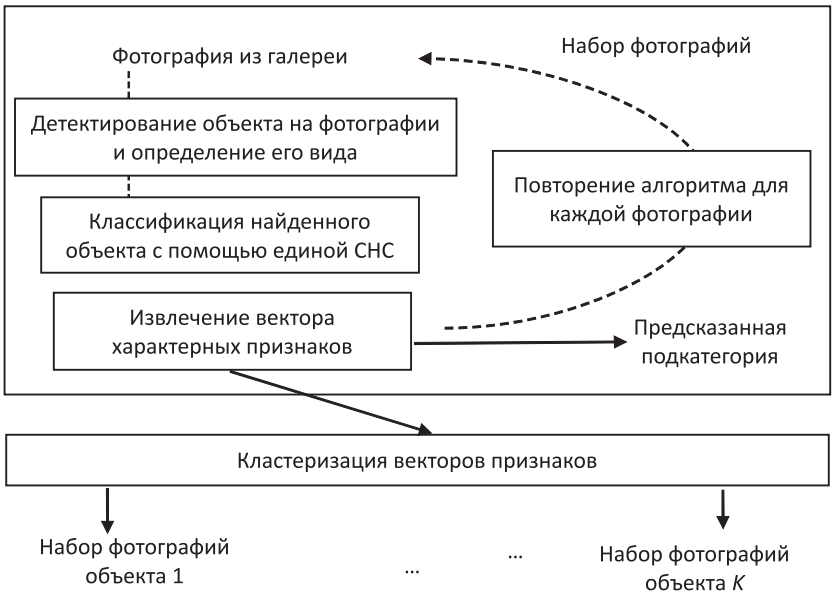

Рис. 3. Схема информационной системы извлечения предпочтений на основе детектирования категорий нескольких различных видов объектов

4. Вычислительный эксперимент

Эксперимент 1: классификации пород домашних животных. Для обучения моделей используются два набора данных (рис. 5, см. третью сторону обложки): Stanford Dogs Dataset [5] и Oxford-IIIT-Pet [6]. Набор Stanford Dogs Dataset содержит изображения $C_1 = 120$ пород собак (150...200 изображений для каждой породы). Oxford-IIIT-Pet содержит породы кошек и собак, для работы были взяты только изображения $C_2 = 12$ пород кошек (примерно 200 изображений на каждую породу).

При подготовке к обучению 70 наугад выбранных изображений каждой категории были помещены в обучающую выборку, а оставшиеся изображения были использованы для тестирования. Применялись различные архитектуры глубоких СНС — как высокоточные вычислительно сложные модели (Inception V3, Inception ResNet v2, ResNet-50, ResNet-101), так и легковесные нейронные сети MobileNet v1, MobileNet v2. Для практического исследования из библиотеки Keras 2.2.5 были взяты СНС, предварительно обученные для классификации изображений из базы данных ImageNet-1000.

В эксперименте сопоставляли два описанных выше способа организации многозадачной СНС: объединение всех подкатегорий (см. рис. 1) для разных видов животных (кошек и собак с общим числом 132 класса) и наличие двух выходов (см. рис. 2) — по одному для каждого вида животных. Каждый классификатор обучался в течение 120 эпох с помощью оптимизатора Adam, при этом в течение первых 20 эпох обучались только веса, связанные с последним слоем (классификатор), а веса базовой сети для извлечения признаков оставались фиксированными.

Для проведения экспериментов был использован ПК с Nvidia GeForce GTX 1080 Ti GPU (12 Гбайт), AMD Ryzen Threadripper 1920X ЦПУ (2.2 ГГц), 64 Гбайт ОЗУ. В табл. 1 и 2 приведены результаты для разных архитектур и двух вариантов многозадачной СНС.

Из сравнения результатов видно, что несмотря на одинаковое число параметров и время классификации (табл. 2), модели с общим выходом (см. рис. 1) в среднем оказались на 1...3 % менее точными по сравнению с моделями с двумя классификаторами (см. рис. 2). Наличие нескольких выходов может считаться встроенной в архитектуру регуляризацией, позволяющей исключить влияние объектов одного вида на выходы, которые ответственны за другие подкатегории [2]. В результате подтверждается вывод о более высоком качестве многозадачных нейронных сетей [15, 16]: выходные слои (см. рис. 2) для собак и кошек обучаются отдельно, что позволяет более качественно настроить параметры для миноритарных классов (в данном случае, пород кошек) и, как следствие, понизить вероятность их ошибочной классификации. Наилучшую точность (90,6 %) показала архитектура Inception v3. При этом "легковесная" MobileNet v2 ($\alpha = 1{,}4$) показывает приемлемую точность, сравнимую с традиционной ResNet-50. По результатам качественного визуального анализа результатов классификации замечено, что ошибки допускаются либо при плохом качестве объекта-животного на изображении, либо для похожих пород.

Для сравнения в табл. 3 приведены известные наилучшие результаты для распознавания пород собак из набора данных Stanford Dogs. Как видно, предложенный подход с многозадачной СНС Inception оказывается на 0,6 %

Таблица 1

Точность классификации пород животных для многозадачных сверточных нейронных сетей

Базовая СНС	СНС с объединением всех подкатегорий		СНС с $N = 2$ выходными слоями	
	Собаки	Кошки	Собаки	Кошки
Inception ResNet v2	0,899	0,815	0,9	0,874
ResNet-50	0,869	0,809	0,859	0,879
ResNet-101	0,872	0,855	0,878	0,884
Inception v3	0,911	0,8	0,906	0,865
MobileNet v2 ($\alpha = 1.0$)	0,788	0,755	0,818	0,84
MobileNet v2 ($\alpha = 1.4$)	0,832	0,844	0,851	0,883

Таблица 2

Размер модели и среднее время классификации одного изображения

Базовая СНС	Число весов, млн	Время классификации, мс
Inception ResNet v2	54,75	10,1
ResNet-50	23,87	6,8
ResNet-101	43,00	9,8
Inception v3	22,55	9,0
MobileNet v2 ($\alpha = 1.0$)	2,48	6,1
MobileNet v2 ($\alpha = 1.4$)	4,62	6,9

Таблица 3

Результаты наилучших известных методов классификации пород собак

Модель	Число изображений каждой породы в обучающем множестве	Точность
Inception ResNet v2 [17]	100	0,900
ResNet-101 [17]	100	0,869
Inception v3 [17]	100	0,889
ResNet50 [18]	100	0,838
ResNet50-CURL [18]	100	0,816
MobileNetV2 [18]	100	0,789
MobileNetV2-CURL [18]	100	0,747
Вероятностная нейронная сеть с проекционными оценками [19, 20]	10	0,729

точнее по сравнению с лучшим известным методом [15].

Эксперимент 2: кластеризация фотографий домашних животных. **В** качестве материала для сравнения алгоритмов кластеризации был собран специальный набор [21] из 190 фотографий двух кошек черного и рыжего цвета (около 40 изображений каждой) и трех собак (рис. 6, см. третью сторону обложки). Большая часть фотографий собак принадлежит одной собаке (колли), при этом около половины ее фотографий сделаны в значительно младшем возрасте, поэтому эти фотографии изначально размечены как две отдельных собаки. Третья собака — черный лабрадор — присутствует примерно на 10 фотографиях. Кроме того, встречаются другие собаки, которые помещены в отдельный кластер выбросов.

В табл. 4 приведены число выделенных кластеров K и значения метрик оценки качества кластеризации в сравнении с реальным распределением по кластерам: Adjusted Rand Index (ARI) и Adjusted Mutual Information (AMI). Использовалась реализация методов кластеризации из библиотек scikit-learn и HDBSCAN. Указаны наилучшие комбинации параметров, вид животного (кошки и собаки группировались отдельно), а также используемый вектор признаков для кластеризации. В параметрах иерархической кластеризации приведен тип межкластерного расстояния, при этом во всех случаях наилучшее качество группировки достигалось для метрики L_1.

Здесь для кошек оптимальное число кластеров — два, а для собак корректными можно считать значения от 3 до 6. Значения ARI, AMI равны 1 при идеальной кластеризации, значения 0,5...0,6 указывают на получение приблизительно верных кластеров. Таким образом, в результате проведенных экспериментов было показано, что кластеризация вектора признаков **x**, извлеченных базовой СНС, может группировать фотографии домашних животных. В то же время для практического применения необходим тщательный выбор параметров для большого обучающего множества.

Заключение

В целом можно сделать заключение, что предложенный подход позволяет осуществить

Таблица 4

Сравнительный анализ методов кластеризации животных

Вид животного	Метод	Параметры	K	ARI	AMI
Собаки	Иерархическая кластеризация	Вектор признаков **x**, Ward	4	0,64	0,55
		Выходы СНС **p**, Average linkage	4	0,641	0,45
	DBSCAN	Вектор признаков **x**, eps = 9, core = 3	4	0,696	0,546
		Выходы СНС **p**, eps = 0.6, core = 3	4	0,549	0,418
	HDBSCAN	Вектор признаков **x**, minPts = 3	5	0,56	0,56
Кошки	Иерархическая кластеризация	Выходы СНС **p**, Complete linkage	2	0,9	0,845
	DBSCAN	Вектор признаков **x**, eps = 9, core = 4	2	1	1
		Выходы СНС **p**, eps = 0.5, core = 5	2	1	1
	HDBSCAN	Вектор признаков **x**, minPts = 3	2	1	1

высокоточное детектирование категорий нескольких различных видов объектов, для которых в обучающем множестве отсутствуют данные о положении на фотографии (обрамляющие прямоугольники). Показано, что для снижения затрат памяти можно использовать многозадачные нейронные сети (см. рис. 2) с несколькими выходами. Экспериментально показано, что такой подход позволяет повысить точность классификации подклассов (пород) домашних животных по сравнению с известными аналогами на основе специализированных нейронных сетей (см. табл. 3). Показано, что обученная нами многозадачная сеть позволяет извлекать характерные признаки объектов, приемлемые для группировки фотографий, содержащих одинаковые объекты (табл. 4).

Основным ограничением предлагаемого подхода является использование идентичных характерных признаков для разных видов объектов. Если для некоторых видов (например, изображений животных) такой подход является приемлемым, то для существенно различающихся объектов наилучшая точность достигается с использованием собственных специализированных СНС. Поэтому в будущих исследованиях необходимо модифицировать многозадачную СНС так, чтобы извлекать характерные признаки на нескольких различных слоях. Выходы первых слоев обычно в достаточной степени независимы от предметной области, поэтому могут быть использованы для классификации совершенно разных видов объектов, но при этом требуют больших объемов обучающих данных. Классификаторы выходов последних слоев могут обучаться даже на малых выборках наблюдений за счет использования доменной адаптации и технологии переноса знаний [2].

Список литературы

1. **Гречихин И. С., Савченко А. В.** Метод анализа предпочтений пользователя по фото- и видеоизображениям на мобильном устройстве на основе нейросетевых детекторов объектов на изображениях // Информационные технологии. 2019. Т. 25. № 9. С. 538—544

2. **Goodfellow I., Bengio Y., Courville A.** Deep Learning (Adaptive Computation and Machine Learning series) // Cambridge, USA, MIT Press, 2016. 800 p.

3. **Ren S., He K., Girshick R., Sun J.** Faster R-CNN: Towards real-time object detection with region proposal networks // Advances in neural information processing systems (NIPS). 2015. P. 91—99.

4. **Savchenko A. V.** Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output ConvNet // PeerJ Computer Science. 2019. Vol. 5, p. 197.

5. **Khosla A., Jayadevaprakash N., Yao B., Li F.-F.** Novel dataset for fine-grained image categorization: Stanford dogs // Proceedings of the CVPR Workshop on Fine-Grained Visual Categorization (FGVC). 2011. Vol. 2.

6. **Parkhi O., Vedaldi A., Zisserman A., Jawahar C.** Cats and dogs // Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2012. P. 3498—3505.

7. **Grechikhin I., Savchenko A. V.** User modeling on mobile device based on facial clustering and object detection in photos and videos // Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA). Springer. 2019. P. 429—440.

8. **Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L. C.** MobilenetV2: Inverted residuals and linear bottlenecks // Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2018. P. 4510—4520.

9. **Lin T. Y., Goyal P., Girshick R., He K., Dollár P.** Focal loss for dense object detection // Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2017. P. 2980—2988.

10. **Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z.** Rethinking the Inception architecture for computer vision // Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2016. P. 2818—2826.

11. **Kolesnikov A., Beyer L., Zhai X., Puigcerver J., Yung J., Gelly S., Houlsby N.** Big Transfer (BiT): General Visual Representation Learning // arXiv preprint arXiv: 1912.11370. 2019 (дата доступа 29.04.2020).

12. **Theodoridis S., Koutroumbas K.** Pattern Recognition, 4th Ed. 2009. 984 p.

13. **McInnes L., Healy J., Astels S.** hdbscan: Hierarchical density based clustering // Journal of Open Source Software. 2017. Vol. 2. N. 11. doi:10.21105/joss.00205.

14. **Разработанное** Android-приложение. URL: https://drive.google.com/open?id=1rThhcKReOb5A9LBIH6jkP8tTiYjoVNWH (дата доступа 29.04.2020).

15. **Liu S., Johns E., Davison A. J.** End-to-end multi-task learning with attention // Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2019. P. 1871—1880.

16. **Yan C., Zhou L., Wan Y.** A multi-task learning model for better representation of clothing images // IEEE Access. 2019. Vol. 7. P. 34499—34507.

17. **Eshratifar A. E., Eigen D., Gormish M., Pedram M.** Coarse2Fine: a two-stage training method for fine-grained visual classification // arXiv preprint arXiv: 1909.02680. 2019 (дата доступа 29.04.2020).

18. **Luo J.-H., Wu J.** Neural network pruning with residual-connections and limited-data // arXiv preprint arXiv:1911.08114. 2019 (дата доступа 29.04.2020).

19. **Savchenko A. V.** Probabilistic neural network with complex exponential activation functions in image recognition // IEEE Transactions on Neural Networks and Learning Systems. 2020. Vol. 31, Iss. 2. P. 651—660

20. **Савченко А. В.** Тригонометрическая система функций в проекционных оценках плотности вероятности нейросетевых признаков изображений // Компьютерная оптика. 2018. Т. 42, № 1. С. 149—158.

21. **Набор** изображений для тестирования кластеризации домашних животных, URL: https://drive.google.com/drive/folders/1-tNB_GR2LkCBsNKQkxB—9ertdllgD7h (дата доступа 29.04.2020)

A. V. Savchenko, Doctor of Sciences, Professor, e-mail: avsavchenko@hse.ru,
I. S. Grechikhin, Postgraduate Student, Senior Lecturer, e-mail: igrechikhin@hse.ru,
National Research University Higher School of Economics, Nizhny Novgorod, Russian Federation

Detection of Specialized Object Categories in Photos from Mobile Device Based on a Multi-Task Neural Network

In this paper we consider the task of user preferences analysis for recommender engines based on a gallery of his or her mobile device. In particular, we propose the novel three-phase method for simultaneous image-based detection and recognition of particular objects. Conventional object detection techniques cannot be applied if there are many categories of the same object (pet breeds, car models, etc.) and there is a lack of large dataset with known bounding boxes for each object category. In order to deal with this issue, we estimate the borders of base objects (dogs, cats, cars, etc.) by using such existing neural network architectures as high precision Faster R-CNN or fast single-shot detectors. Secondly, the visual features (embeddings) of each object are extracted by using a multi-task convolutional neural network model with several outputs — one for each type of object. Finally, these embeddings are used to predict the concrete categories and group different photos of the same object by using cluster analysis techniques. The proposed approach is implemented in a special mobile application for Android. Experimental results for recognizing dog and cat breeds are presented. It is demonstrated that our method makes it possible to improve the accuracy of dog detection and recognition when compared to the known single-task neural nets. Moreover, we gather a special dataset of real photos with pets to estimate the clustering quality. It is shown that the L_1-normed features extracted by our multi-task model may be grouped rather accurately if hierarchical agglomerative clustering or HDBSCAN method are used.

Keywords: *image processing, convolutional neural networks, mobile systems, pet breed recognition, hierarchical clustering, multi-task learning, object detection*

Acknowledgments. The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE University) in 2019-2020 (grant No. 19-04-004) and by the Russian Academic Excellence Project "5-100"

DOI: 10.17587/it.26.586-593

References

1. **Grechikhin I., Savchenko A. V.** Analysis of user preferences using photos and videos from mobile device based on object detection and neural networks, *Informacionnye tehnologii*, 2019, vol. 25, no. 9, pp. 538—544 (in Russian).

2. **Goodfellow I., Bengio Y., Courville A.** Deep Learning (Adaptive Computation and Machine Learning series). Cambridge, USA, MIT Press, 2016. 800 p.

3. **Ren S., He K., Girshick R., Sun** J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *Advances in neural information processing systems (NIPS)*, 2015, pp. 91—99.

4. **Savchenko A. V.** Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output ConvNet, *PeerJ Computer Science*, 2019, 5:e197.

5. **Khosla A., Jayadevaprakash N., Yao B., Li F.-F.** Novel dataset for fine-grained image categorization: Stanford dogs, *Proceedings of the CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, 2011, vol. 2.

6. **Parkhi O., Vedaldi A., Zisserman A., Jawahar C.** Cats and dogs, *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2012, pp. 3498—3505.

7. **Grechikhin I., Savchenko A. V.** User modeling on mobile device based on facial clustering and object detection in photos and videos, *Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, Springer, 2019, pp. 429—440.

8. **Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L. C.** MobilenetV2: Inverted residuals and linear bottlenecks, *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2018, pp. 4510—4520.

9. **Lin T. Y., Goyal P., Girshick R., He K., Dollár P.** Focal loss for dense object detection, *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017, pp. 2980—2988.

10. **Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z.** Rethinking the Inception architecture for computer vision, *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016, pp. 2818—2826.

11. **Kolesnikov A., Beyer L., Zhai X., Puigcerver J., Yung J., Gelly S., Houlsby N.** Big Transfer (BiT): General Visual Representation Learning, arXiv preprint arXiv: 1912.11370. 2019 (date of access 29.04.2020).

12. **Theodoridis S., Koutroumbas K.** *Pattern Recognition*, 4th Edition, 2009, 984 p.

13. **McInnes L., Healy J., Astels S.** hdbscan: Hierarchical density based clustering, *Journal of Open Source Software*, 2017, vol. 2, no. 11, doi:10.21105/joss.00205.

14. Developed Android-application, available at: https://drive.google.com/open?id=1rThhcKReOb5A9LBIH6jkP8tTiYjoVNWH (date of access 29.04.2020)

15. **Liu S., Johns E., Davison A. J.** End-to-end multi-task learning with attention, *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2019, pp. 1871—1880.

16. **Yan C., Zhou L., Wan Y.** A multi-task learning model for better representation of clothing images, *IEEE Access*, 2019, vol. 7, pp. 34499—34507.

17. **Eshratifar A. E., Eigen D., Gormish M., Pedram M.** Coarse2Fine: a two-stage training method for fine-grained visual classification, arXiv preprint arXiv: 1909.02680. 2019 (date of access 29.04.2020).

18. **Luo J.-H., Wu J.** Neural network pruning with residual-connections and limited-data, arXiv preprint arXiv:1911.08114. 2019 (date of access 29.04.2020).

19. **Savchenko A. V.** Probabilistic neural network with complex exponential activation functions in image recognition, *IEEE Transactions on Neural Networks and Learning Systems*, 2020, vol. 31, iss. 2, pp. 651—660.

20. **Savchenko A. V.** Trigonometric series in orthogonal expansions for density estimates of deep image features, *Computer Optics*, 2018, vol. 42, no. 1, pp. 149-158 (in Russian).

21. **Dataset** of cats and dogs images for testing of clustering, available at: https://drive.google.com/drive/folders/1-tNB_GR2LkCBsNKQkxB—9ertdllgD7h (date of access 29.04.2020).

Рисунки к статье А. В. Савченко, И. С. Гречихина

«ДЕТЕКТИРОВАНИЕ СПЕЦИАЛИЗИРОВАННЫХ КАТЕГОРИЙ ОБЪЕКТОВ НА ФОТОГРАФИЯХ В МОБИЛЬНЫХ УСТРОЙСТВАХ НА ОСНОВЕ МНОГОЗАДАЧНОЙ НЕЙРОСЕТЕВОЙ МОДЕЛИ»

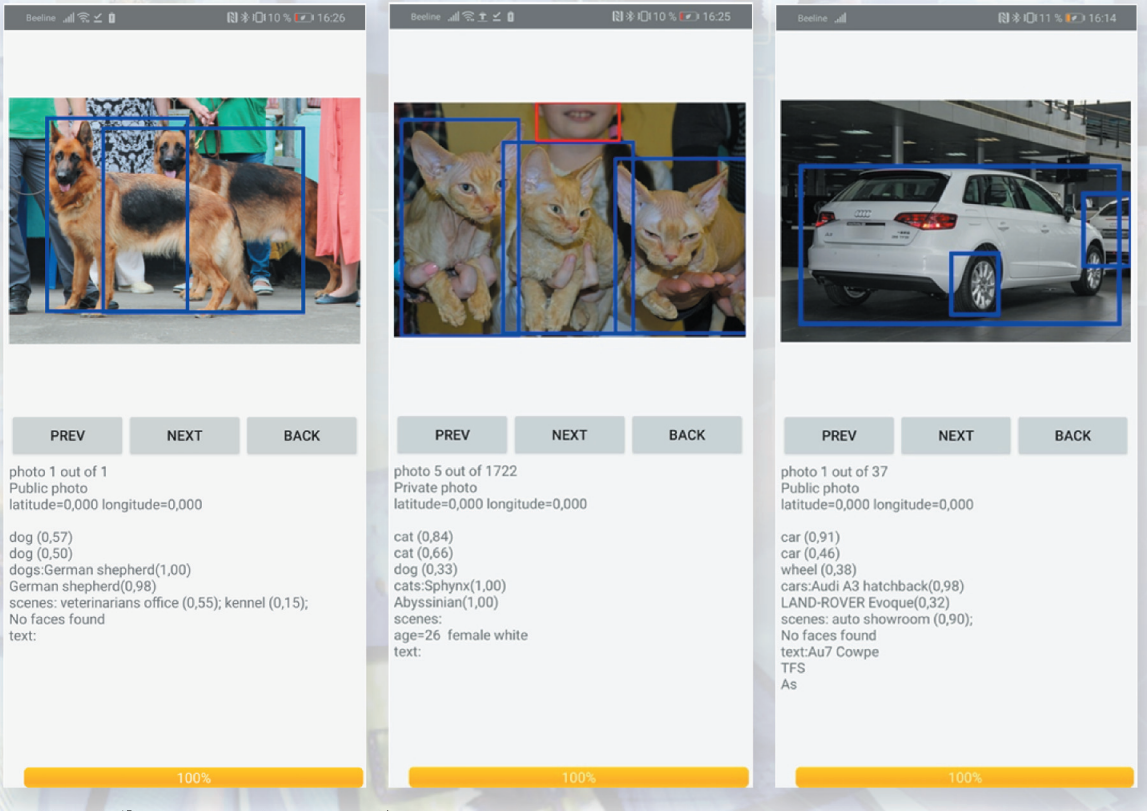

Рис. 4. Экранные формы мобильного приложения, реализующего предложенный подход

Рис. 5. Примеры изображений из наборов данных Stanford Dogs и Oxford-IIIT-Pet

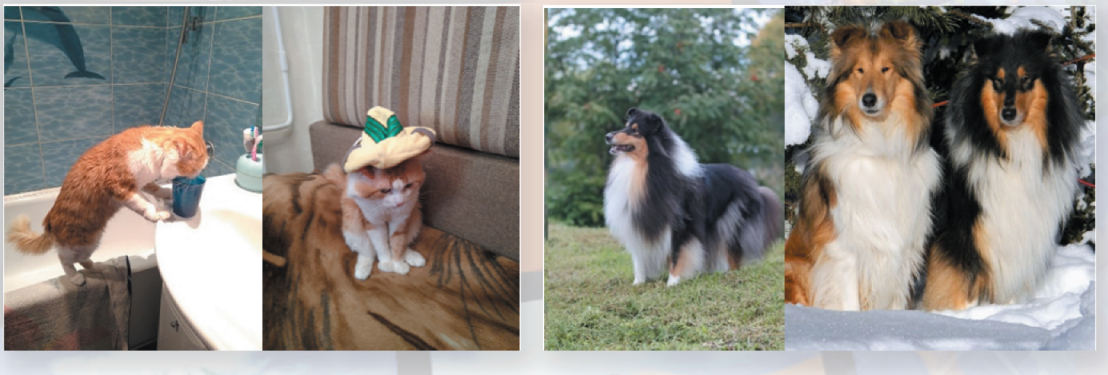

Рис. 6. Примеры изображений из собранного набора фотографий для тестирования качества кластеризации