Федеральное государственное автономное образовательное учреждение
высшего образования «Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук
Основная образовательная программа
Прикладная математика и информатика

ГРУППОВАЯ КУРСОВАЯ РАБОТА

Программный проект на тему

"Классификация пола и возраста людей на фотографии"

Выполнили студенты группы 171, 3 курса,
Биршерт Алексей Дмитриевич,
Шабалин Александр Михайлович

Руководитель КР: старший преподаватель
Соколов Евгений Андреевич

Куратор: Магистр, руководитель группы нейросетевых технологий компьютерного зрения
Овчаренко Сергей Александрович

Москва 2020

Содержание

Аннотация

Автоматическое предсказание пола и возраста человека по фотографиям, полученным в самых разных условиях - это важная и сложная задача, находящая применение во многих областях жизнедеятельности людей. В своем проекте мы стремились воплотить подход к решению этой задачи, основанный на сверточных нейронных сетях. Свое решение мы разбили на две составных части - детектирование лица человека и ключевых точек на его лице и дальнейшее предсказание пола и возраста по признакам лица.

Для детектирования мы используем архитектуру RetinaFace с ResNet-18 в качестве основой модели. В качестве обучающих данных мы используем датасет WIDER FACE с добавленными к нему пятью ключевыми точками для каждого лица. На тестовой выборке наш детектор получает значение метрики Precision равное 83%.

Для классификации найденных лиц мы используем модель на основе двух ResNet-18. В качестве обучающих данных мы испольуем датасеты IMDB-WIKI-101 и FGNET. В качестве тестовых данных мы используем датасет Adience, наша модель получает для возраста значение метрики Accuracy 45% и значение метрики One-off-accuracy 80%, для пола значение метрики Accuracy 85%.

В итоге мы получили рабочую модель, способную обрабатывать большие массивы фотографий, детектировать на них людей и предсказывать для их пол и возраст в автоматическом режиме. Ссылка на гитхаб с проектом - https://github.com/birshert/age_gender_classification.

Ключевые слова—Определение возраста и пола, Детектирование лиц, Компьютерное зрение, Глубокое обучение

Automatically predicting real age and gender from face images acquired in unconstrained conditions is an important and challenging task in many real-world applications. In our project we intended to construct an approach to predicting a person's real age and gender from photograph based on convolutional neural networks. Our solution is divided into solving two separate subtasks - detecting one's face and it's landmarks and further age and gender estimation based on facial features.

For face detection we use RetinaFace architecture with ResNet-18 as a backbone. For training we use WIDER FACE dataset with added five facial landmarks for every face. On testing subset we achieve 83% Precision result.

For real age and gender estimation we use a model based on two ResNet-18.

For training we use IMDB-WIKI-101 dataset and FGNET dataset. For testing we use Adience dataset, our model achieves 45% Accuracy, 80% One-off-accuracy for age estimation and 85% Accuracy for gender prediction.

As a result, we got a working model capable of processing large volumes of photos, detecting people faces on them and predicting their real age and gender in automatic mode. Github project link - https://github.com/birshert/age_gender_classification.

Keywords—Age and gender estimation, Facial detection, Computer vision, Deep learning

1 Введение

1.1 Описание предметной области

Во многих сферах деятельности данные такого рода как пол и возраст человека имеют большое значение. Антропология, маркетинг, социология - во всех этих областях наличие инструмента, позволяющего быстро получать такие данные, очень ценно. Хорошим источником для получения этих данных могут служить фотографии людей. Однако исследование большого объема фотографий вручную занимает слишком много времени, за которое данные могут устареть и перестать быть актуальными, к тому же ручная разметка довольно дорого стоит. В своей работе мы стремимся решить задачу автоматизации получения данных о возрасте и поле человека по его фотографии. Данная задача относится к задачам компьютерного зрения.

Самые ранние подходы к классификации пола и возраста опирались на известные зависимости размеров и форм черт лица, так называемые антропометрические модели. Более поздние подходы начали использовать сверточные нейронные сети. Все методы так или иначе используют в первую очередь изображение лица человека в качестве основного признакового описания. В зависимости от преследуемой цели инженеры ставят перед собой следующие задачи: улучшить качество предсказаний, увеличить скорость работы и научиться работать с фотографиями плохого качества.

1.2 Постановка задачи

Проблему предсказания возраста и пола нужно рассматривать в разделении на подзадачи - детекция и выравнивание лица, определение пола и определение возраста.

Наша цель в задаче детекции и выравнивания лица заключается в наилучшем определении ключевых точек на лице, соответствующих глазам, и наилучшем определении ограничивающего контура, обрамляющем лицо на

фотографии. При этом для нас гораздо предпочтительнее не найти какое-нибудь лицо, чем выделить объект, не являющийся лицом. По этой причине мы будем стараться максимизировать метрику $Precision$:

$$Precision = \frac{|\text{detected faces}|}{|\text{all detected objects}|}$$

Мы будем считать, что найденный объект является лицом, если модель уверена в нем больше, чем на 50%. Также стоит заметить, что правильность выставления ключевых точек для нас тоже не критична, поэтому функционал ошибки мы зададим как

$$L = L_{cls} + 0.25L_{box} + 0.1L_{pts},$$

где L_{cls} - ошибка классификации, L_{box} - ошибка контура и L_{pts} - ошибка ключевых точек.

Наша цель в задаче определения пола заключается в достижении наилучшего качества бинарной классификации.

$$\sum_{i=1}^{n} [\mathcal{G}(x_i) = y_i] \to \max_{\mathcal{G}},$$

где x - объекты, y - их пол, $\mathcal{G}$ - модель, предсказывающая пол.

В задаче определения возраста мы будем минимализировать разницу между предсказанным возрастом и реальным возрастом человека.

$$\sum_{i=1}^{n} \|\mathcal{A}(x_i) - y_i\| \to \min_{\mathcal{A}},$$

где x - объекты, y - их возраст, $\mathcal{A}$ - модель, предсказывающая возраст.

Итогом проектной работы должна стать программно реализованная система, принимающая на вход от пользователя каталог фотографий и возвращающая пол и возраст людей, запечатленных на фотографиях. В первую очередь мы будем стремиться получить наилучшее качество классификации,

но при этом не будем забывать про время работы. Система должна опираться на использование сверточных нейронных сетей.

Дальнейшая работа описана в следующих главах: Обзор литературы, Детектирование лиц, Классификация гендерных и возрастных групп, Описание системы для пользователя, Заключение.

"Детектирование лиц" выполнено Александром Шабалиным, "Классификация гендерных и возрастных групп" Алексеем Биршертом.

2 Обзор литературы

2.1 Архитектура ResNet

Так как при детекции и классификации лиц мы пользуемся архитектурой сверточной сети ResNet, мы считаем нужным пояснить, как она устроена.

ResNet [1] создавалась по подобию VGG [2]. Так же, как у VGG большинство ее сверток используют размер ядра 3x3, при проходе в глубину размер карты признаков уменьшается, а количество каналов увеличивается. Используемая в нашем случае версия ResNet-18 имеет следующую архитектуру. Первый слой является сверткой с ядром размера 7x7 и сдвигом 2, после которого идет 3x3 maxpooling с шагом 2. Затем располагаются 4 блока в каждом блоке содержится по 4 сверточных слоя с ядром 3x3 с функцией активации ReLU. Количество фильтров сверточных слоев в первом блоке равно 64, с каждым следующим блоком число фильтров удваивается. Во всех блоках, кроме первого, первый сверточный слои имеет шаг 2, из-за чего размер карты признаков сокращается вдвое. Таким образом с каждым новым все блоки требуют одинаковое количество операций с плавующей точкой и поэтому их выполнение занимает одинаковое количество времени. В конце располагаются слой average pool и полносвязный слой.

Отличительной чертой ResNet являются быстрые соединения. Их смысл состоит в передачи информации одного слоя не только следующему прямо

после него, но и следующему за ним через два. Такой подход предотвращает затухание градиента и позволяет модели продолжать обучаться. Быстрые соединения повторяются в ResNet через каждые два слоя.

2.2 Детектирование лиц

Детектирование лиц на фотографии - одна из древнейших задач компьютерного зрения, возникшая в 1990-х годах. Первые хорошие результаты появились в 2004 году. Описанный в статье [3] метод находил лица с помощью признаков Хаара, используя каскад детекторов, обученных алгоритмом AdaBoost. В 2014 году в статье [4] был предложен метод, использующий Deformable Parts AgeGender (DPM). Его идея заключается в нахождении зависимостей между подвижными частями. Например, лицо представляется, как нечто, состоящее из глаз, носа, рта, расположенных в некотором антропоморфическом виде. Однако все описанные методы показывали довольно плохие результаты в сложных случаях, так как опирались на ограниченный, придуманный людьми набор признаков. Поэтому методы, основанные на сверточных нейронных сетях быстро вытеснили остальные. Лучшие известные на данный момент подходы описаны в статье [5]. Все из них используют нейронные сети (обычно ResNet) для получения признаков. Так как модель сама находит признаки и зависимости, результат получается лучше.

Немаловажной задачей является и выравнивание лиц. Самый быстрый метод получить выровненное лицо - определить ключевые точки на лице и преобразовать изображение так, чтобы эти точки были на заранее определенных местах. В статье [6] описан метод вычисления координат основных точек на лице человека - окаймляющих лицо, глаза, нос, рот и брови. Вычисление точек происходит с помощью каскада регрессоров, обучаемых с помощью градиентного бустинга.

В нашей работе мы пользуемся архитектурой RetinaFace [7]. На данный момент она является state-of-the-art в задаче детектирования лиц и показы-

вает впечатляющие результаты. Одна из ее отличительных черт - умение предсказывать пять ключевых точек, необходимых для выравнивания изображения.

2.3 Классификация возраста и пола

В прошлом, такие задачи решались с помощью алгоритмов, построенных на основе исследования фиксированных, выделенных людьми, признаков, но это не приносило удовлетворительных результатов. Так происходило потому что люди, запечатленные на фотографиях в неформальных условиях, имеют слишком большую вариацию поз, размеров относительно размера кадра, освещения и так далее.

Все известные подходы к предсказанию возраста и пола человека по фото опираются на исследование его лица. Самые ранние подходы к определению возраста людей по фотографии использовали в качестве ключевого фактора различия в пропорциях и размерах черт лица в зависимости от возраста и пола - антропометрические модели [8]. Все они вычисляли положение ключевых точек на лице - глаза, нос, рот, скулы - и анализировали их в дальнейшем.

Более поздние подходы начали применять сверточные нейронные сети. Сверточные сети используются в этой задаче в силу своей способности к извлечению признаков из изображений. А как мы упомянули ранее, именно невозможность сконструировать признаки служила препятствием для качественного анализа. Так, в статье [9] описано использование сверточных нейронных сетей небольшой глубины - три сверточных слоя и два полносвязных. В статье описана работа с датасетом Adience [10], в котором возраст представлен в виде 8 возрастных групп. Таким образом модель училась предсказывать наиболее вероятный диапазон возраста. У авторов статьи удалось добиться результата в $50.7\% \pm 5.1\%$ Accuracy, $84.7\% \pm 2.2\%$ One-off-accuracy для предсказания возрастной группы, $86.8\% \pm 1.4\%$ Accuracy для классификации пола на Adience benchmark, где One-off-accuracy это метрика, равная доле ответов,

которые попали либо в правильную возрастную группу, либо в соседнюю.

С развитием области компьютерного зрения и появлением новых, более сложных, позволяющих достигать всё лучшие результаты, архитектур сверточных нейронных сетей, на основе этих архитектур появляются и решения для задачи предсказания возраста и пола. В более современных статьях, например [11], используется глубокая сверточная нейронная сеть архитектуры VGG-16, что позволяет улучшить качество по сравнению с предшествующими решениями. Стоит отметить, что в этой статье описано предсказание реального возраста от 0 до 100 лет включительно, каждая категория это целое число от 0 до 100. То есть модель училась предсказывать более точный возраст, нежели диапазон. В этой статье описан процесс сбора и подготовки данных для обучения на датасете IMDB-WIKI-101[12]. В результате авторы статьи добиваются результата MAE в 3.22 в конкурсе [13].

В статье [14] используется архитектура RoR, Residual networks of Residual networks, которая представляет собой модификацию моделей ResNet. Авторы статьи достигают качества в $67.3\% \pm 3.6\%$ Accuracy, $97.5\% \pm 0.7\%$ One-off-accuracy для предсказания возрастной группы на Adience benchmark.

Хочется отметить наличие статей, использующих архитектуры на основе ResNet или RoR с LSTM (архитектура реккурентной нейронной сети) [15]. В этой работе сверточные нейронные сети используются для выделения глобальных признаков, отвечающих за возраст, а реккурентные нейронные сети используются для выделения и работы с локальными признаками, что в итоге позволяет добиться впечатляющих результатов. Авторы приводят результат в $67.8\% \pm 3\%$ Accuracy, $97.6\% \pm 0.6\%$ One-off-accuracy для предсказания возрастной группы на Adience benchmark.

3 Детектирование лиц

3.1 Описание выбранного метода

Для детектирования лиц была использована архитектура нейронной сети RetinaFace[7]. Она состоит из двух частей: основной модели и пирамиды признаков. Основная модель - это сверточная нейронная сеть, предназначенная для выявления признаков из изображения. В качестве основной модели может выступать MobileNet [16], VGG, ResNet и другие.

Идея пирамиды признаков довольно проста. При детектировании лиц мы хотим находить лица разных размеров. Для этого строятся две пирамиды из пяти карт признаков в каждой. Первая получается проходом снизу в верх, а вторая - сверху вниз. Слоями первой пирамиды являются выходы четырех слоев основной модели такие, что размер каждого следующего выхода в два раза меньше размера предыдущего. Так мы получаем карты разного размера, что позволяет детектировать как маленькие, так и большие лица. Вторая пирамида необходима из-за того, что на ранних слоях сверточной сети содержится значительно меньше семантической информации, а значит, предсказания лиц на ранних слоях менее точны и мы должны как-то компенсировать это. Слои второй пирамиды мы получаем, проходя сверху вниз следующим образом. Первый (верхний) слой получается с помощью наложения свертки с ядром размера 1x1 на верхний слой первой пирамиды. Каждый следующий из пяти слоев получается путем поэлементого суммирования предыдущего слоя, увеличенного в два раза, со сверткой размера 1x1 соответствующего слоя первой пирамиды. Таким образом нам удается передать большим по размеру картам признаков семантическую информацию меньших карт, компенсировав ее нехватку. Подобная передача информации от первой пирамиды также моделирует обходные связи, используемые в ResNet и улучшающие обучение модели.

Важным объектом данной архитектуры являются якори. При детектировании лиц на карте признаков карта разбивается на сетку из регионов, каж-

дый регион содержит в себе k прямоугольников различных форм и размеров (якорей). Каждый якорь несет в себе предсказание об ограничивающем контуре, ключевых точках и бинарной классификации лицо/не лицо в заданном им прямоугольнике, то есть якорь содержит $4 + 10 + 2 = 16$ параметров. Предсказания получаются с помощью применения нескольких сверточных слоев с размерами ядер 5x5 и 7x7 к картам признаков второй пирамиды и конкатенированием результатов этих сверток. Число выходных каналов этих сверток должно равняться 16, так как именно столько парамертов обозначают предсказание лица.

В нашей реализации в качестве основной модели используется ResNet18 предобученный на ImageNet-1000 [17]. Выходы четырех блоков ResNet являются четыремя картами признаков первой пирамиды, пятая карта признаков получается наложением свертки с ядром размера 3x3 и шагом 2 на последнюю карту признаков. Так как датасет ImageNet содержит в себе картинки с различными изображениями, а в нашей задаче необходимо находить лица людей, мы замораживем только первый слой ResNet, а остальные дообучаем. Выбор ResNet с 18 слоями обосновывается тем, что в нашей задаче также важна скорость работы. С увеличением размера сверточной сети время ее работы заметно увеличивается, а качество растет не так сильно.

3.2 Подготовка данных и обучение

Для обучения модели был использован датасет WIDER FACE [18]. Он содержит 32,203 фотографии с 393,703 лицами на них. Для каждого лица хранятся координаты пяти ключевый точек: две для глаз, одна для носа и две для рта, а также координаты прямоугольной рамки, ограничивающей лицо. В качестве аугментации мы пробовали делать случайный переворот изображения, однако он не улучшил результаты, поэтому мы решили от него отказаться. Входные изображения масштабируются до квадратных следую-

щим образом: большая сторона становится равной 256, а к меньшей добавляется нулевой отступ с двух сторон так, чтобы размер итогового изображения составлял 256 на 256. При обучении был использован Adam оптимизатор со скоростью обучения 10^{-3}. Модель обучается в течение 20-ти эпох с размером батча 32.

Ошибка модели считается по формуле:

$$L = L_{cls}(p, p^*) + 0.25L_{box}(t, t^*) + 0.1L_{pts}(l, l^*),$$

где p - вероятность лица, p^* - истинный ответ (0 или 1), для подсчета ошибки классификатора $L_{cls}(p, p^*)$ используется софтмакс ошибка для бинарных классов классов (лицо/не лицо).
$t = \{t_1, t_2, t_3, t_4\}$ и $t^* = \{t_1^*, t_2^*, t_3^*, t_4^*\}$ предсказанные и истинные координаты ограничивающей рамки соответственно. Для подсчета ошибки $L_{box}(t, t^*)$ используется функция Smooth-L_1 loss.
$l = \{l_{x1}, l_{y1}, \ldots, l_{x5}, l_{y5}\}$ и $l = \{l_{x1}^*, l_{y1}^*, \ldots, l_{x5}^*, l_{y5}^*\}$ предсказанные и истинные координаты ключевых точек лица. Для подсчета ошибки $L_{pts}(l, l^*)$ так же используется Smooth-L_1 loss.

3.3 Выравнивание

Перед передачей найденных лиц классификатору возраста и пола необходимо их выровнить. Выравнивание производится на основе двух ключевых точек для глаз, найденных вместе с лицами. Для этого находится угол отклонения прямой, проведенной с помощью точек глаз, от горизонтали. После этого мы домножаем матрицу изображения на матрицу поворота таким образом, что линия глаз становится горизонтальной. Мы не используем остальные точки, так как они избыточны, а в некоторых случаях даже мешают. Например, расположение точек рта у улыбающегося человека и у серьезного отличаются, но это отличие не должно никак влиять на выравнивание. Глаза же всегда располагаются на одном месте относительно лица, что позволяет

их считать хорошей опорой для выравнивания.

3.4 Эксперименты

Для определения с выбором основной модели были проведены сравнения между ResNet18, ResNet50, VGG и MobileNet. По результатам этого эксперимента выяснилось, что ResNet50 получает лучший результат, но для предсказания ответа ему требуется гораздо больше времени, чем ResNet18. MobileNet работает быстрее остальных, но ошибка такой модели оказывается больше остальных. VGG показывает хороший результат, но работает медленнее всех. В результате выбор остановился на ResNet18, потому что баланс скорости и качества детектирования лиц этой модели мы считаем оптимальным.

Для ускорения времени работы алгоритма мы пробовали уменьшать число карт признаков в пирамиде признаков, однако в таком случае некоторые слишком большие или слишком маленькие лица переставали обнаруживаться, что крайне негативно отражалось на результатах модели.

3.5 Результаты

Полученная модель хорошо справляется с поставленной задачей, получая значение метрики Precision равное 83%. Для нашей задачи эта метрика имеет крайне важное значение, так как ты не должны классифицировать несуществующих людей. Несмотря на то, что реализация авторов RetinaFace получала 91%, мы считаем это хорошим результатом, так как из-за нехватики вычислительных мощностей пришлось почти в 3 раза снизить размер стороны входного изображения, использовать ResNet18 вместо ResNet152, а также уменьшить количество эпох обучения, что заметно сказывается на результатах работы.

3.6 Примеры работы модели

(a) Модель смогла найти 50 лиц из 61 указанного в ответе.

(b) Модель хорошо справилась со своей задачей, найдя все повернутые к камере лица, однако она также выделила пустую часть стола.

4 Классификация гендерных и возрастных групп

Эта часть выполнена Алексеем Биршертом.

4.1 Описание выбранного метода

Для предсказания пола и возраста используются две нейронные сети, состоящие из основной и выходной моделей каждая. В качестве основной модели используется глубокая нейронная сеть ResNet-18, без последнего полносвязного слоя. В качестве выходной модели используется персептрон из двух полносвязных слоёв с нелинейностью ReLU и дропаутом между ними. На вход основной модели подаются изображения 227 на 227 пикселей, 3 канала цвета - R, G, B, выход основной модели подаётся на вход выходной модели. Выходом нейросети будет выход выходной модели. Первая нейронная сеть предназначена для классификации пола и имеет в выходной модели 512 и 256 входных и выходных нейронов в первом слое, 256 и 2 во втором соответственно. Вторая нейронная сеть предназначена для классификации возраста и имеет в выходной модели 512 и 512 входных и выходных нейронов в первом слое, 512 и 101 во втором соответственно. На вход подаются фотографии лиц людей, выделенные и выровненные с помощью модели детектирования лиц. Предсказанный пол определяется как номер выходного нейрона соответствующей нейронной сети с максимальным значением - первый это "женский", второй "мужской". Предсказанный возраст определяется следующим образом: сначала для вектора значений выходных нейронов соответсвующей нейронной сети применяется преобразование софтмакс, затем значения умножаются на соответствующий им возраст. После вектор суммируется - получаем матожидание возраста при вероятностном распределении,

выданном моделью.

$$AGE = \sum_{i=0}^{100} i \cdot softmax(x)_i, \quad softmax(x)_i = \frac{\exp(x_i)}{\sum_{j=0}^{100} \exp(x_j)}.$$

4.2 Описание и подготовка данных

В качестве датасета для обучения двух вышеописанных моделей были избраны датасеты IMDB-WIKI-101 [12] и FGNET [20]. Оба датасета имеют в мета-данных метки пола "женщина"или "мужчина" и имеют метки возраста в виде целых чисел от 0 до 100 включительно. Распределение возраста в датасете IMDB-WIKI-101 имеет вид нормальной кривой со средним около 35 лет, имея малое количество объектов с возрастом меньше 10 лет или больше 90. Для увеличения количества данных с возрастом до 10 лет был избран датасет FGNET, в котором большая часть объектов это дети до 15 лет. Для улучшения сходимости нейронных сетей была произведена предобработка всех объектов - в итоговую выборку не были включены следующие объекты: объекты с плохо различимыми лицами (показатель уверенности модели распознавания лиц в том, что это лицо, ниже фиксированного значения), объекты с некорректно заполненными данными по полу/возрасту, объекты со слишком маленькими фотографиями. На каждом изображении было выделено и выровнено лицо. Отступ от границы лица был поставлен на 40%, чтобы лицо целиком оказывалось на выделенной зоне. Итого было получено около 200 тысяч объектов, которые были в дальнейшем поделены с сохранением баланса классов 1 к 19 на валидационную и обучающую выборки соответственно.

В качестве датасета для тестирования был избран датасет Adience [10], по которому известно большое количество результатов различных моделей. Из него были исключены объекты с некорректным описанием пола или возраста. Итого было получено почти 11 тысяч объектов для тестовой выборки. В Adience метки возраста в формате 8 групп - 0: [0, 2], 1: [4, 6], 2: [8, 12], 3: [15,

20], 4: [25, 32], 5: [38, 43], 6: [48, 53], 7: [60, 100]. В связи с этим, необходимо было решить как относить к этим группам метки реального возраста от 0 до 100. Было принято решение относить к той группе, граница которой ближе по модулю, в случае равенства относить к первой в порядке следования.

4.3 Постановка задач обучения

Задача классификации пола является задачей бинарной классификации, целевая переменная для одного объекта это число 0 или 1. Для обучения была выбрана перекрестная энтропия - функция ошибки со следующей формулой:

$$loss(x, y) = -x_y + \log\left(\sum_{j=1}^{K} \exp(x_j)\right),$$

где x - вектор значений выходных нейронов нейронной сети, y - целевая переменная. Для оценки качества классификации использовались метрика доля правильных ответов, так как выборки сбалансированны по полу.

Задача классификации возраста является задачей многоклассовой классификации. Если смотреть на задачу предсказания возраста по фотографии с точки зрения человека, человек гораздо точнее способен угадать диапазон возраста, нежели точный возраст. Поэтому задача классификации возраста была интерпретирована как задача с множественными правильными ответами - каждому объекту может соответствовать набор правильных классов. Целевой переменной для одного объекта служил вектор из нулей и единиц, единицы на позициях правильных классов. Правильные классы определялись как значения возраста, которые отличаются по модулю от правильного не больше, чем на фиксированное число, которое было гиперпараметром (далее об этом в разделе эксперименты 4.4). Для обучения была выбрана бинарная перекрестная энтропия -

$$loss(x, y) = sum(L), \quad L = \{l_1, \dots l_N\}, \quad l_i = \left(y_i \log(x_i) + (1 - y_i) \log(1 - x_i)\right),$$

где x это вектор значений выходных нейронов нейронной сети после применения сигмоидного преобразования, y - целевая переменная. Для оценки качества классификации использовались средний модуль отклонения (далее MAE) и доля объектов, у которых MAE не превышает фиксированного числа (далее CS-5).

4.4 Эксперименты

В качестве основных было рассмотрено два варианта - нейронная сеть на основе статьи [14] и нейронная сеть на основе статьи [15]. Однако стоит заметить, что анализ и использование модели из второй статьи на порядок сложнее, так как помимо сверточных нейронных сетей там используется блок LSTM, для работы с которым необходимо иметь соответствующий опыт. Поэтому было принято решение базировать работу на основе первой статьи, слегка упростив архитектуру.

Необходимо было решить вопрос выбора архитектуры базовой модели. Всего было опробовано три различных архитектуры - MobileNet, ShuffleNet [19], ResNet. Лучше всего себя проявила архитектура ResNet-18. MobileNet и ShuffleNet незначительно быстрее, не смотря на реализацию, однако достаточно сильно проигрывали в качестве. В итоге была выбрана архитектура ResNet.

Так как ResNet-18 предобучен на датасете ImageNet-1000, он имеет очень хорошую способность выделять признаки из изображения. Известно, что первые (входные) сверточные слои обученных моделей очень похожи между собой даже для различных задач. Таким образом, было решено веса первого сверточного слоя ResNet-18 взять из предобученной на ImageNet-1000 модели и заморозить в процессе обучения, дообучая веса всех остальных слоев.

Вторым необходимо было решить вопрос выбора количества моделей - одна нейронная сеть из общей базовой модели и двух паралелльных выходных моделей или две отдельные нейронные сети из основной и выходной модели. Однако стоит заметить, что для обеих задач нейронная сеть и должна вы-

делять признаки лица, признаки, отвечающие за пол, несколько отличаются от признаков, которые соответствуют возрасту. Также в силу использования функций ошибки, которые имеют разный масштаб и решают частично разные задачи на моменте основной модели, и сложности в подборе гиперпараметров для комбинирования этих функций ошибки, нейронная сеть из одной общей основной модели очень плохо сходилась и достигала намного меньшего качества во всех проведенных экспериментах. В итоге было принято решение использовать две паралелльные нейронные сети, по одной на решаемую задачу.

В процессе подбора гиперпараметров для обучения были выбраны следующие значения: темп обучения был выставлен на $1e-3$ для первых 40 эпох, далее $1e-4$ для модели предсказания возраста, $1e-4$ для первых 30 эпох и $2e-5$ далее для модели предсказания пола; коэффициент $L2$ регуляризации был установлен на $1e-3$ для обеих моделей; размер батча был установлен в силу ограничений на память ГПУ и в качестве дополнительной регуляризации при обучении на 128; размер входных картинок 227 на 227 пикселей; дропаут 0.1 перед 4 слоём ResNet, 0.2 перед первым полносвязным слоём и 0.4 между слоями. В качестве размера окна для правильного возраста после сравнений было выбрано число в 5 лет. Модели, обученные с такими гиперпараметрами показали наилучшие результаты.

В процессе выбора аугментации были выбраны следующие трансформации - в процессе обучения при проходе по датасету каждое изображение преобразовывалось к квадрату 256 на 256 пикселей, затем из него выбирался случайный квадрат со стороной 227 пикселей, который с вероятностью 1/2 мог быть отражен вдоль вертикальной оси, проходящей через его центр. Во время тестирования каждое изображение преобразовывалось к квадрату 256 на 256 пикселей, затем из него по центру вырезался квадрат со стороной 227 пикселей.

4.5 Результаты

Как итог модели показали следующие показатели метрик на обучающей и валидационной выборках: для возраста MAE 5 и 5.6 соответственно, CS-5 0.71 и 0.69 соответственно, для пола Accuracy 0.94 и 0.93 соответственно.

Модель показала следующие результаты на датасете Adience: для возраста Exact-accuracy $49.1\% \pm 4.1\%$ и One-off-accuracy $84.4\% \pm 3.6\%$, для пола значение метрики Accuracy $83.7\% \pm 2.7\%$. Значения метрик получаются подсчетом соответствующих метрик для каждой из 8 возрастных групп и вычисления среднего арифметического от посчитанных показателей.

Всего было написано более 2300 строк кода на языке программирования Python3, все модели были написаны и обучены с использованием фреймворка машинного обучения PyTorch.

5 Описание системы для пользователя

6 Заключение

Решая поставленную задачу о классификации людей на фотографиях по полу и возрасту мы получили алгоритм, позволяющий быстро и качественно находить лица людей на фотографиях, а затем предсказывать их пол и возраст. При решении мы опробовали множество различных подходов и остановились на архитектуре RetinaFace для детектирования лиц и двух ResNet для классификации пола и возраста.

Наша модель для детекции лиц показывает значение метрики Precision равное 83% на датасете WIDER FACE [18]. Наша модель для классификации пола и возраста показывает значение метрики Accuracy 45% и значение метрики One-off-accuracy 80%, для пола значение метрики Accuracy 85% на датасете Adience [10].

Для улучшения качества классификации в дальнейшем у нас есть несколько идей, которые мы не успели реализовать. Определенно положительно бы сказалась дальнейшая работа по улучшению качества всех моделей, по уменьшению их размера и скорости работы. Также имеет смысл добавить в структуру нашего решения фильтрацию неживых лиц - памятников и лиц на рекламных щитах и постерах.

7 Список литературы

Список литературы

1. Deep Residual Learning for Image Recognition. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. In IEEE 2015

2. Very deep convolutional networks for large-scale image recognition. Karen Simonyan, Andrew Zisserman. In ICLR 2015

3. Robust Real-Time Face Detection. Paul Viola, Michael J. Jones. In IEEE 2003

4. Face detection without bells and whistles. Makrus Mathias, Rodrigo Beneson, Marco Pedersoli, Lus Van Gool. In ECCV 2014

5. Accurate Face Detection for High Performance. Faen Zhang, Xinyu Fan, Guo Ai, Jianfei Song, Yongqiang Qin, Jiahong Wu. In ArXiv 2019

6. One Millisecond Face Alignment with an Ensemble of Regression Trees. Vahid Kazemi and Josephine Sullivan. In IEEE 2014

7. RetinaFace: Single-stage Dense Face Localisation in the Wild. Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, Stefanos Zafeiriou. In ArXiv 2019

8. Age Classification from Facial Images. Young H. Kwon, Niels da Vitoria Lobo. In IEEE 1994

9. Age and Gender Classification using Convolutional Neural Networks. Gil Levi, Tal Hassner. In IEEE 2015

10. Adience Database

11. DEX: Deep EXpectation of apparent age from a single image. Rasmus Rothe, Radu Timofte, Luc Van Gool. In IJCV 2016

12. IMDB-WIKI-101 Database

13. ChaLearn Looking at People 2015: apparent age and cultural event recognition datasets and results. Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baro, Jordi Gonzalez, Hugo J. Escalante, Marc Oliu, Dusan Misevic, Ulrich Steiner, Isabelle Guyon. In EFP 2015

14. Age group and gender estimation in the wild with deep RoR architecture. Ke Zhang, Ce Gao, Liru Guo, Miao Sun, Xingfang Yuan, Tony X. Han, Zhenbing Zhao and Baogang Li. In IEEE 2017

15. Fine-grained age estimation in the wild with attention LSTM networks. Ke Zhang, Na Liu, Xingfang Yuan, Xinyao Guo, Ce Gao, Zhenbing Zhao and Zhanyu Ma. In ArXiv 2018

16. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam. In ArXiv 2017

17. ImageNet-1000 Database

18. WIDER-FACE Database

19. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, Jian Sun. In ECCV 2018

20. FGNET Database