Ссылка на статью:
// Ученые записки УлГУ. Сер. Математика и информационные технологии. 2022, № 2, с. 123-129.

Поступила: 21.11.2022
Окончательный вариант: 08.12.2022

УДК 519.68

Трансферное обучение на основе архитектуры нейронной сети EfficientNetV2B0 в задаче построения 3D маски лица человека

Шабалин А. С.**, *Елисеев И. М.

*alexshabalin73@gmail.com
УлГУ, Ульяновск, Россия

В работе применяется подход трансферного обучения для дообучения архитектуры искусственной нейронной сети EfficientNetV2B0 в задаче создания 3D маски человека. Разработанная модель прогнозирует 478 точек лица человека, каждая из которых представлена в виде вектора из трех координат. Представлен способ разметки данных и подход к обучению.

***Ключевые слова**: искусственные нейронные сети, трансферное обучение, 3D маска человек, EfficientNetV2B0.*

Введение

В последние годы мобильные устройства стали основным источником фотографий для «непрофессионалов», что в свою очередь увеличивает спрос на качественные снимки [1]. Большой упор идет на вычислительную фотографию (computational photography). Практически ни одна современная фотография не обходится без применения различных алгоритмов улучшения качества, которые сильно зависят от графов вычислений (искусственных нейронных сетей) и глубокого обучения. Вычислительной мощности современных смартфонов хватает не только на обработку изображения, но и обработку потока видео в реальном времени, тут одна из решаемых задач - оценивание 3D ориентиров (точек) лица в режиме реального времени [2]. Благодаря этому возможно так популярное использование масок и фильтров.

Актуальной остается проблема улучшения точности разметки лица в неидеальных условиях: качество камер смартфонов, качество освещения, различные ракурсы и т.п. Для решения такой задачи можно использовать нейросетевые модели, которые можно обучить предсказывать не только двухмерное расположение точек в пространстве, но и трехмерную маску лица, имея в качестве входных данных лишь двухмерную фотографию.

В данной статье описывается подход обработки двухмерной фотографии лица человека для получения 478 точек трехмерного пространства, с целью дальнейшего построения 3D модели лица. Подход основан на трансферном обучении (Transfer Learning) и архитектуре искусственной нейронной сети EfficientNetV2 [4].

Разметка данных

Для обучающей выборки использовались первые десять тысяч изображений из набора данных Flickr-Faces-HQ Dataset (FFHQ) (https://github.com/NVlabs/ffhq-dataset). Выбор ограниченного числа изображений обусловлен вычислительной производительностью системы, на которой запускался алгоритм. Изображения состоят из лиц людей и случайного фона. Лица на изображениях могут быть случайно повернуты вплоть до 90 градусов. Никакого дополнительного признакового описания для данного набора данных не проводилось. Все изображения имеют размерность 1024x1024x3 и сохранены в png формате.

Для разметки данных использовалась библиотека MediaPipe Face Mesh [3], которая содержит различные решения для обнаружения лица на изображении, а также четырёхстах семидесяти восьми ориентиров – трехмерных векторов координат точек, для определения лица, глаз, губ и других частей лица человека, из которых 468 довольно точно определяют положение глаз, носа, губ и других частей лица человека (маска представлена на рис. 1), а также 10 дополнительных точек (по 5 на каждый глаз) для определения зрачка. Координаты x и y нормализуются в зависимости от размера кадра. Значение третьей координаты не отражает расстояние от камеры до лица, а взаимосвязано с другими аналогичными координатами, что может быть достаточно для вычисления угла поворота

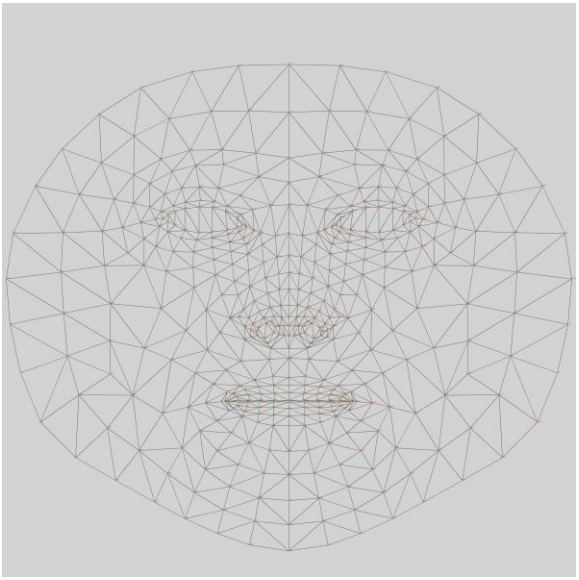

Рис 1. Каноничное изображение для разметки точек лица

Пример исходного и размеченного изображения представлен на рис. 2.

Рис.2. Пример одного из изображений FFHQ (слева) и его разметка 478 точками

Обработка данных и обучение

Оптимизационные методы для уменьшения времени работы нейронной сети (инференса), обрезка (pruning), TTA, аугментация, частичная выборка данных не применялись.

Набор данных для обучающей выборки представляет собой тензоры, полученные методами библиотеки MediaPipe Face Mesh. Исходное изображение преобразуется в тензор цветного изображения размерностью 256x256x3 со значениями в диапазоне [0...1].

Выходными данными является матрица значений координат точек размерностью 3x478 и значениями в диапазоне [0...1], координаты определяются относительно верхнего левого угла (0, 0). Для наложения точек применяется также предобработка изображения для сжатия к размерности входных данных и перевода значений субпикселей от [0...255] к размерности [0...1].

Для обучения модели применяется предобученная версия EfficientNetV2 - EfficientNetV2B0, на наборе данных ImageNet, который содержит порядка 50 миллионов размеченных изображений и это один из самых часто используемых наборов данных для обучения и валидации новых архитектур нейронных сетей.

Изначально применяемая модель была обучена для решения задачи классификации, для того чтобы адаптировать модель под требуемую задачу определения матрицы координат точек, применена техника переноса знаний (transfer learning). Во многих архитектурах нейронных сетей начальные слои изучают общую информацию, а слои на последнем уровне более специфичные признаки. Например, первые слои могут запоминать текстуру, цвет, общую картину, а последние глаза, рот, родинки и т.д. Таким образом исходная модель EfficientNetV2B0 остается в целом без изменений, за исключением добавления новых признаков и обучения последних слоев с целью перепрофилировать модель под поставленные задачи.

Для тренировки модели использовались следующие гиперпараметры: размер пакета обучения (batch_size) 32, стартовое значение скорости обучения (initial learning rate) 0.01, динамически изменяется по ходу обучения.

Выбор архитектуры сети

EfficientNetV2 является логическим продолжением первой версии EfficientNet, семейства неросетей специализирующихся на скорости тренировки и инференса. EfficientNetV2B0 является уменьшенной версией стандартной архитектуры EfficientNetV2-S данного семейства, содержит порядка 7.4 миллиона параметров и лучшей долей правильных ответов в задачи классификации изображений равной 78.7%. В таблице 1 представлена стандартная архитектура EfficientNetV2-S.

Таблица 1. Архитектура EfficientNetV2-S

Этап	Оператор	Шаг	Каналы	Слои
0	Conv3x3	2	24	1
1	Fused-MBConv1, k3x3	1	24	2
2	Fused-MBConv4, k3x3	2	48	4
3	Fused-MBConv4, k3x3	2	64	4
4	MBConv4, k3x3, SE0.25	2	128	6
5	MBConv6, k3x3, SE0.25	1	160	9
6	MBConv6, k3x3, SE0.25	2	256	15
7	Conv1x1 & Pooling & FC	-	1280	1

Данная модель показывает хороший баланс между временем обработки данных и достигаемой точностью. Сравнение скорости обучения нейронной сети, а также доли правильных ответов приведено в [3], мы в свою очередь приводим иллюстрацию из данной статьи показанную на рис. 3.

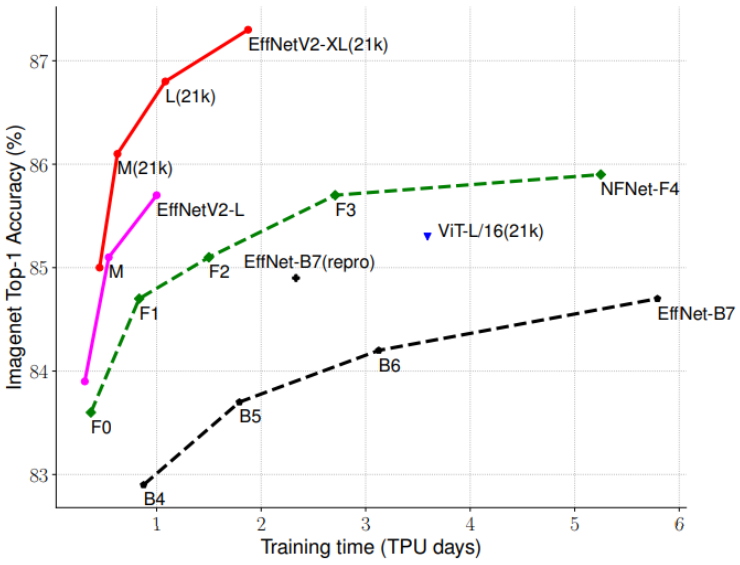

Рис. 3. Сравнение время обучения и точности различных архитектур нейронных сетей

Вместо 7 этапа начальной архитектуры (таблица 1) использовалась собственная реализация, более подходящая под решаемую задачу. Применяется BatchNormalization для повышения производительности и стабилизации работы нейронной сети, разделимая свертка с функцией активации ReLU, flatten для выравнивая входных данных и сохранения неизменного размера пакета данных и выходной полносвязный слой размера 3x478 с функцией активации сигмоиды.

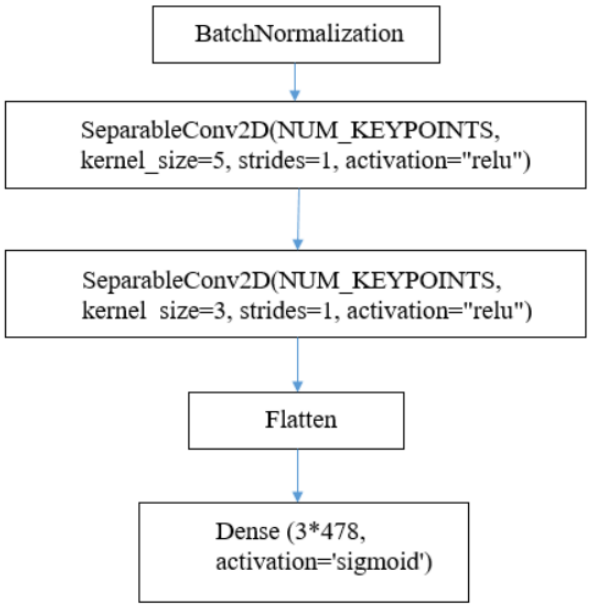

Рис. 4. Схема собственных добавленных слоев в архитектуру EfficientNetV2-S

Функционал качества

Для определения качества модели в данной задачи использовалась метрика средней квадратичной ошибки (MSE), вычисляющее Евклидово расстояние между двумя точками – предсказанными координатами некоторой точки и ее истинным значением. С учетом того, что все данные масштабированы в диапазон [0…1], данная метрика может отражать длину относительно изображения, на которую модель ошибается в своих предсказаниях. Среднее значение ошибки на тестовой выборке 8.1070e-04.

Выводы и результаты

Применение методов трансферного обучения является эффективным инструментом обучения моделей искусственных нейронных сетей, которые позволяют использовать более сложные архитектуры, обученные на миллионах параметров, что в свою очередь не представляется возможным на обычных стационарных компьютерах. Таким образом, в процессе переобучения изменяются параметры только нескольких последних слоев, что позволяет перепрофилировать алгоритм под класс решаемых задач разметки лица человека. На рис. 5 представлена 3D маска лица человека, представленного на рис. 2.

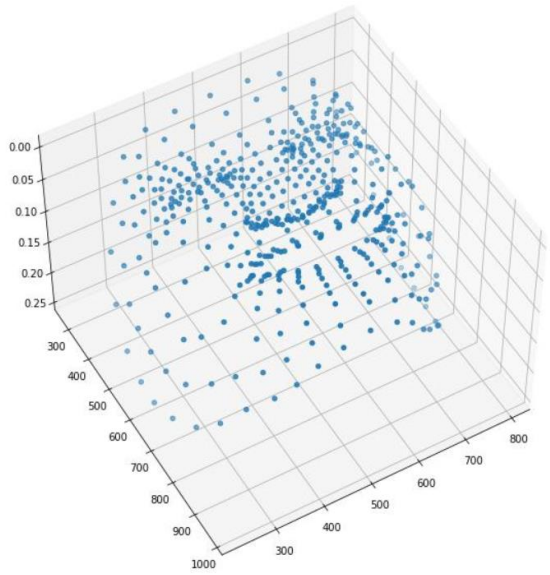

Рис. 5. Пример полученной 3D маски

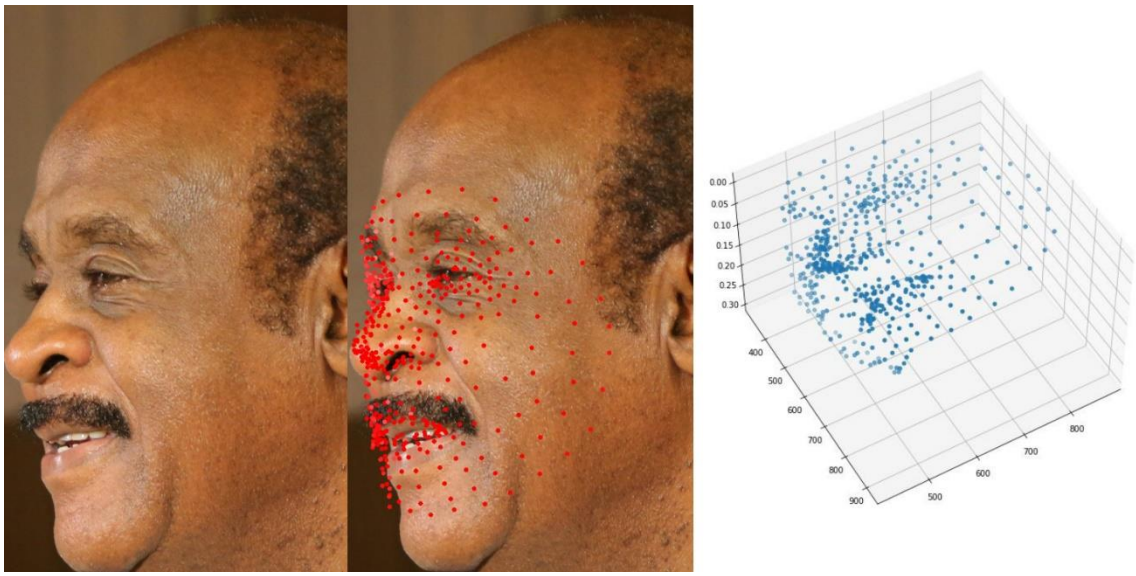

Рис. 6. Пример наложения точек на лицо с поворотом

Разработанная модель может быть использована для наложения различных масок на лицо человека, отслеживания поворота головы (рис. 6), перемещения в пространстве относительно камеры и оцифровки структуры лица. Разработанная модель представлена в открытом доступе по ссылке https://github.com/EmpyEmpt/face-mesh.

Список литературы

1. Ignatov A. et al. Dslr-quality photos on mobile devices with deep convolutional networks // *Proceedings of the IEEE International Conference on Computer Vision*. 2017. С. 3277-3285.
2. Pan J. et al. Deep mesh reconstruction from single rgb images via topology modification networks //*Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019. С. 9964-9973.

3. Radmehr A., Asgari M., Masouleh M. T. Experimental Study on the Imitation of the Human Neck-and-Eye Pose Using the 3-DOF Agile Eye Parallel Robot Based on a Deep Neural Network Approach //arXiv preprint: arXiv:2111.00452, 2021.
4. Tan M., Le Q. Efficientnetv2: Smaller models and faster training // *International Conference on Machine Learning*. PMLR, 2021. С. 10096-10106.

Transfer learning based on the EfficientNetV2B0 neural network architecture in the problem of constructing a 3D human facemask

Shabalin, A. S. [*]***, Eliseev, I. M.***

[*]alexshabalin73@gmail.com
Ulyanovsk State University, Ulyanovsk, Russia

The paper uses the transfer learning approach to retrain the EfficientNetV2B0 artificial neural network architecture in the problem of creating a 3D human mask. The developed model predicts 478 points of a person's face, each of the points is represented as a vector of three coordinates. A method of data labeling and an approach to learning are presented.

Keywords: *artificial neural networks, transfer learning, 3D human mask, EfficientNetV2B0.*