РЕГРЕССИЯ 2D КЛЮЧЕВЫХ ТОЧЕК ЛИЦА НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ

А.В. Тылецкий

Белорусский государственный университет, г. Минск;

a.tyletsky@gmail.com;

науч. рук. – Д. И. Пирштук, ст. преп.

В данной работе мы сначала поставим задачу нахождения ключевых точек лица и опишем способ оценки точности решения задачи. После чего рассмотрим процесс подготовки и аугментации данных используемого датасета, опишем общий алгоритм обучения нейронных сетей для задачи регресии ключевых точек лица, произведём сравнение различных архитектур нейросетей, как классических, таких как ResNet и MobileNetV2, так и более современных – EfficientNet, MobileNetV3, MobileViT, на поставленной задаче, покажем влияние различных аспектов в общем алгоритме обучения нейронных сетей.

Ключевые слова: регрессия точек лица; нейронные сети; ResNet; MobileNetV2; EfficientNet; MobileNetV3; MobileViT; функция потерь Wing; аугментация; набор данных LaPa.

ВВЕДЕНИЕ

Распознавание ключевых точек лица используется для виртуальной реконструкции лица, распознавания эмоций, отслеживание взгляда водителя для мониторинга внимания, бьютификации лиц и многих других практических областях. Нейронные сети продемонстрировали поразительное улучшение качества в решении данной задачи и в настоящее время изучаются многими специалистами в этой области.

ПОСТАНОВКА ЗАДАЧИ

Обозначим через I входное изображение. Пусть также $x_i \in R^2$ есть координаты i-ой ключевой точки изображения I. Тогда через вектор $S = (x_1^T, x_2^T, \ldots, x_p^T)^T \in R^{2p}$ можно обозначить все p ключевых точек лица на изображении I. Задача обнаружения ключевых точек лица состоит в том, чтобы найти такую функцию $f : I \to S$, которая по входному изображению I предсказывает вектор ключевых точек лица S. Количество точек лица p, а также точное отображение i-ой точки лица в ее координаты на картинке заданы в датасете.

ОЦЕНКА ТОЧНОСТИ

Наиболее популярной метрикой в решении задачи регресии ключевых точек лица является NME (Normalized Mean Error). Нормализация в нашем случае будет происходить на расстояние d между зрачками.

$$NME = \frac{1}{K}\sum_{k=1}^{K} NME_k = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{p}\sum_{i=1}^{p}\frac{\| \breve{x}_i - x_i \|_2}{d_k},$$

где K – количество изображений, d_k – расстояние между зрачками на k-ом изображении. Существуют и другие метрики оценки точности, подробнее о которых написано в [1].

ПОДГОТОВКА И АУГМЕНТАЦИИ ДАННЫХ

В работе использовался набор данных LaPa [2], созданный с целью повышения точности и качества обучения. Датасет содержит 22176 изображений разнообразных лиц, где для каждого изображения размечены 106 ключевых точек лица. В обучении использовались заранее выделенные авторами части датасета: тренировочные данные (18176 изображений), валидационные данные (2000 изображений) и тестовые данные (2000 изображений).

В данной статье будут рассматриваться предобученные нейронные сети на задаче ImageNet. В связи с этим связаны стандартные для этого действия по предобработке изображений: перевод значений пикселей в диапазон [0, 1], масштабирование изображения до размеров 224×224, нормализация изображений. Также для каждой ключевой точки P мы применим преобразование $(P-(112,112))/112$, которое в большинстве случаев переведет значение координат точки P в диапазон [-1, 1].

На вход в нейросеть будем подавать не всю картинку, а только достаточно маленькую область, содержащую лицо, ключевые точки которого нужно найти. В процессе обучения для получения ограничивающей лицо рамки мы будем случайно равновероятно выбирать один из двух способов: обрезка лица по истинным ключевым точкам или использование уже обученного детектора лиц. Первый способ является классическим примером того, как можно подсматривать в целевые данные, однако его совместное со вторым способом использование на этапе обучения может помочь ускорить и улучшить качество обучения. На этапе валидации и тестирования мы будем использовать только второй способ.

Для обучения нейросетей будут использоваться следующие достаточно простые и тем не менее эффективные аугментации:

- Случайный поворот изображения (на угол до 35°).

- Манипуляции с цветом: изменение яркости, контрастности, насыщенности и оттенка изображения.
- Случайный сдвиг границ ограничивающей лицо рамки. Каждую из 4 границ мы будем равновероятно сдвигать на [-5%,+5%] от текущего положения.

РЕЗУЛЬТАТЫ ОБУЧЕНИЯ. СРАВНЕНИЕ АРХИТЕКТУР

Для решения задачи мы рассмотрим как классические архитектуры нейросетей, такие как ResNet и MobileNetV2, так и более современные – это EfficientNet, MobileNetV3, MobileViT.

Инициализировать нейросети будем весами, предобученными на задаче ImageNet. Также будем использовать разогрев для шага обучения первые 3 эпохи. Количество выходных нейронов в последнем слое в каждой из архитектур поставим равным 214, что есть удвоенное количество ключевых точек лица. Будем решать задачу регрессии используя среднеквадратичную функцию потерь. Обучение нейросетей производилось на языке Python, используя библиотеку PyTorch.

Итоговые результаты метрики NME на тестовой выборке в зависимости от различных параметров отображены на рисунке 1.

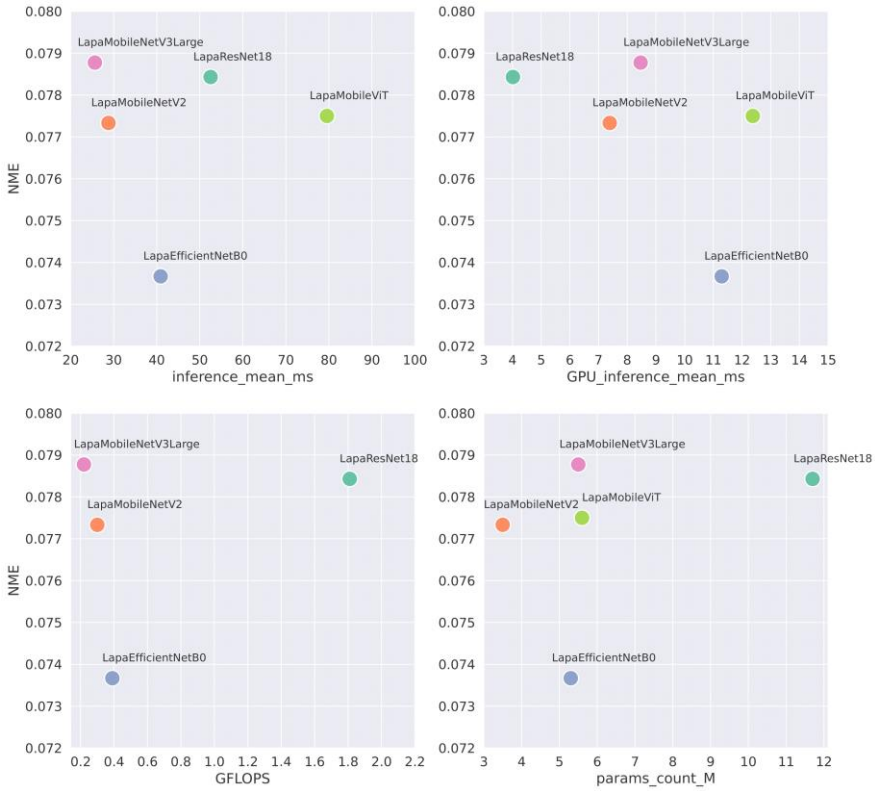

Рис. 1. Зависимость NME на тестовой выборке от различных параметров: времени инференса на ЦПУ, на ГПУ, GFLOPS, количества обучаемых параметров.

Все архитектуры, кроме EfficientNetB0 дали примерно одинаковый показатель метрики. Противоречивое время инференса ResNet18 объясняется тем, что все остальные рассмотренные архитектуры больше оптимизированы для применения на мобильных устройствах. MobileViT ожидаемо оказался медленнее остальных архитектур из-за наличия трансформерных компонент.

ВЛИЯНИЕ РАЗОГРЕВА, ПРЕДОБУЧЕНИЯ, АУГМЕНТАЦИЙ, ФУНКЦИИ ПОТЕРЬ

Выберем в качестве архитектуры MobileNetV3 и рассмотрим некоторые модификации процесса обучения и подготовки данных. Результаты представлены в таблице 1. Все рассмотренные модификации, кроме использования функции потерь Wing [3], в той или иной мере ухудшают итоговое качество модели.

Таблица 1

MobileNetV3 с различными модификациями

Тип модификации	NME
MobileNetV3-Large без модификаций	0,0788
Функция потерь Wing	**0,0536**
Без разогрева	0,0807
Замороженные слои (первые 3 из 16 блоков)	0,0817
Без предобучения (случайно-инициализированные веса)	0,0881
Только детектор для получения рамки лица во время обучения	0,0884
Без аугментаций поворота и манипуляций с цветом	0,0823
Предыдущее + без сдвига рамки	0,0931

Функция потерь Wing менее чувствительна к выбросам и гораздо более чувствительна к средним и малым ошибкам, что улучшает обучение в целом. Обучив архитектуру EfficientNetB0, использую Wing, можно получить еще более точную модель, которая на тестовой части выборки добивается показателя метрики в 0,0524.

Библиографические ссылки

1. Khabarlak K., Koriashkina L. Fast Facial Landmark Detection and Applications: A Survey // Journal of Computer Science and Technology. – 2022. – Т. 22. – №. 1.
2. Liu Y. et al. A new dataset and boundary-attention semantic segmentation for face parsing // Proceedings of the AAAI Conference on Artificial Intelligence. – 2020. – Т. 34. – №. 07. – С. 11637-11644.
3. Feng Z. H. et al. Wing loss for robust facial landmark localisation with convolutional neural networks // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2018. – С. 2235-2245.