Записки научных
семинаров ПОМИ
Том 499, 2021 г.

А. В. Савченко

ВЫЧИСЛИТЕЛЬНО ЭФФЕКТИВНЫЕ АЛГОРИТМЫ КЛАССИФИКАЦИИ ИЗОБРАЖЕНИЙ НА ОСНОВЕ ПОСЛЕДОВАТЕЛЬНОГО АНАЛИЗА

§1. Введение

Задача математической классификации изображений состоит в том, чтобы поставить в соответствие поступающему на вход изображению X один из $C > 1$ классов (категорий). При обучении с учителем каждый c-й класс задается с помощью $N(c) \geqslant 1$ обучающих примеров $X_{c;n}, n \in \{1, \ldots, N(c)\}$. Такая задача возникает во многих интеллектуальных системах [12], например, при распознавании объектов на видео, при распознавании атрибутов (пола, возраста, национальности, эмоций) лиц, а также в рекомендательных системах, предсказывающих интересы пользователя по набору фотографий в галерее мобильного устройства. Для их решения обычно используются методы обработки векторов признаков, извлеченных из изображений с помощью сверточных нейронных сетей (СНС) [6]. В таком случае входному X и эталонным $X_{c;n}$ изображениям ставятся в соответствие D-мерные векторы признаков $\mathbf{x} = [x^{(1)}, \ldots, x^{(D)}]$ и $\mathbf{x}_{c;n} = [x^{(1)}_{c;n}, \ldots, x^{(D)}_{c;n}]$. Зачастую накладываются дополнительные ограничения на реализацию систем с использованием типовых мобильных устройств, поэтому применяемые модели должны целиком помещаться в оперативную память устройства и не использовать чрезмерно сложные в вычислительном плане алгоритмы и нейросетевые архитектуры. Проблема высокой вычислительной сложности современных методов классификации изображений особенно усиливается для высоких размерностей признакового пространства $D \gg 1$ и большого числа классов $C \gg 1$ (сотни альтернатив) [12].

Ключевые слова: распознавание изображений, последовательный анализ, классификация атрибутов лиц, классификация эмоций, распознавание расы, сверточная нейронная сеть.

Исследование выполнено за счет гранта Российского научного фонда (проект No. 20-71-10010).

Таким образом, цель настоящей статьи состоит в повышении вычислительной эффективности алгоритмов классификации изображений. Предложено использовать методы статистического последовательного анализа [22] для обработки главных компонент нейросетевых векторов признаков и досрочного прекращения прямого прохода (inference) при получении достаточно надежного решения на ранних слоях СНС. Рассматривается возможность последовательного обучения эффективной нейронной сети для сходных задач классификации атрибутов лиц. Полученные результаты и сделанные по ним выводы рассчитаны на широкий круг специалистов в области компьютерного зрения.

§2. Последовательный анализ в задачах классификации изображений

2.1. Статистическая классификация векторов признаков высокой размерности. Для декорреляции компонент векторов признаков оценивается матрица преобразования Φ размерности $D \times \widetilde{D}$ с помощью анализа главных компонент обучающей выборки $\{\mathbf{x}_{c;n}\}$, где $\tilde{D} \leqslant \min(D, N)$. В результате векторы $\mathbf{x}_n$ преобразуются в $\widetilde{D}$-мерные векторы линейно независимых компонент $\widetilde{\mathbf{x}}_n = [\widetilde{x}_{c;n}^{(1)}, \ldots, \widetilde{x}_{c;n}^{(\tilde{D})}]$. Аналогично, дескриптору входного изображения $\mathbf{x}$ ставится в соответствие вектор главных компонент $\tilde{\mathbf{x}} = [\tilde{x}^{(1)}, \ldots, \tilde{x}^{(\tilde{D})}]$.

Для повышения вычислительной эффективности вместо анализа всех главных компонент воспользуемся последовательным анализом [23]. Разобьем вектор главных компонент на L непересекающихся частей, так, что l-я часть ($l \in \{1, \ldots, L\}$) состоит из $m = L/\tilde{D}$ компонент с номерами $d_{l-1} + 1, \ldots, d_l$, где $d_l = \min(lm, \tilde{D})$ [15]. Если рассматривать задачу распознавания изображений в терминах статистической классификации, можно предположить, что последовательность $\{(\mathbf{x}_{c;n}, c) | \, c \in \{1, \ldots, C\}, n \in \{1, \ldots, N(c)\}\}$ является выборкой независимых одинаково распределенных пар случайных величин. Тогда задача сводится к проверке C гипотез о распределении вектора признаков $\mathbf{x}$, а ее оптимальное байесовское решение может быть записано в виде

$$c_l^* = \operatorname*{argmax}_{c \in C_l} \frac{N(c)}{N} \widehat{f}_c(\widetilde{\mathbf{x}}(l)), \tag{1}$$

где на первом шаге множество потенциальных решений C_l содержит все классы ($C_1 = \{1, \ldots, C\}$), а $\widehat{f}_c(\widetilde{\mathbf{x}}(l))$ – оценка плотности вероятности признаков входного изображения на l-м шаге на основе его первых d_l главных компонент $\widetilde{\mathbf{x}}(l)$. В выражении (1) предполагается, что априорные вероятности каждого класса оцениваются как отношение числа обучающих примеров $N(c)$ этого класса к общему размеру обучающей выборки $N = \sum_{c=1}^{C} N(c)$.

В рамках последовательного анализа на каждом шаге изменяется множество классов C_l так, чтобы включать только достаточно надежные классы [15], полученные, например, с помощью сравнения отношения правдоподобия с некоторым наперед заданным порогом δ [22]:

$$C_{l+1} = \left\{ c \in C_l \,\middle|\, \frac{\widehat{f}_c(\widetilde{\mathbf{x}}(l))}{\widehat{f}_{c_l^*}(\widetilde{\mathbf{x}}(l))} \geqslant \delta \right\}. \tag{2}$$

Если множество $C_{l+1} = \{c_l^*\}$ содержит только один максимально правдоподобный класс, процесс поиска на этом шаге завершается. В результате вычислительная сложность принятия решений существенно снижается, так как на l-м шаге необходимо оценивать плотности вероятности не всех C классов, а только категорий из множества C_{l+1}.

Оценка распределения c-го класса в (1) может быть получена, например, с помощью непараметрических методов вида

$$\widehat{f}_c(\widetilde{\mathbf{x}}(l)) = \frac{1}{N(c)} \sum_{n=1}^{N(c)} K(\widetilde{\mathbf{x}}(l), \widetilde{\mathbf{x}}_{c;n}(l)), \tag{3}$$

где $K(\widetilde{\mathbf{x}}(l), \widetilde{\mathbf{x}}_{c;n}(l))$ – некоторое ядро. Если воспользоваться ядерной функцией Розенблатта-Парзена, критерий (1), (3) представляет собой реализацию вероятностной нейронной сети (ВНС) [20]. Ее вычислительная сложность линейно зависит от числа обучающих примеров N и размерности вектора признаков lm, что существенно ограничивает применимость ВНС на практике. Традиционный способ преодоления указанного недостатка за счет обработки только малого числа главных компонент зачастую приводит к значимому снижению точности классификации. В то же время стоит отметить, что при использовании последовательного анализа можно сохранять квадраты L_2-расстояний между векторами $\widetilde{\mathbf{x}}(l)$ и $\widetilde{\mathbf{x}}_{c;n}(l)$, использующихся при вычислении ядра

Розенблатта-Парзена, и, тем самым, снизить вычислительную сложность, т.к. на каждом шаге потребуется сравнить только m следующих компонент.

Для дальнейшего повышения вычислительной эффективности предположим, что все главные компоненты статистически независимы:

$$\log \widehat{f}_c(\widetilde{\mathbf{x}}(l)) = \sum_{d=1}^{d_l} \log \widehat{f}_{c;J}(\widetilde{x}^{(d)}). \tag{4}$$

Это выражение можно эффективно вычислить на основе оценки плотности $\widehat{f}_c(\widetilde{\mathbf{x}}(l-1))$ на предыдущем шаге в рамках последовательного анализа (1)-(2):

$$\log \widehat{f}_c(\widetilde{\mathbf{x}}(l)) = \log \widehat{f}_c(\widetilde{\mathbf{x}}(l-1)) + \sum_{d=d_{l-1}+1}^{d_l} \log \widehat{f}_{c;J}(\widetilde{x}^{(d)}), \tag{5}$$

где $f_c(\widetilde{\mathbf{x}}(0)) = 1$ и $d_0 = 0$.

2.2. Проекционные оценки плотности вероятности классов. Для оценки плотности вероятности заменим ядро Розенблатта-Парзена на ортогональные ядерные функции. Так как на практике обычно применяется L_2-нормализация вектора признаков, каждая его компонента принимает значение из ограниченного диапазона $[-1; 1]$, следует использовать разложение неизвестной плотности вероятности d-й главной компоненты с помощью тригонометрического ряда Фурье. Как известно, такая оценка может приводить к отрицательным значениям оценки плотности вероятности [16]. Воспользуемся модификацией ВНС с проекционными оценками плотности вероятности главных компонент на основе ядра Фейера:

$$\widehat{f}_{c;J}(\widetilde{x}^{(d)}) = 0.5 + \sum_{j=1}^{J} \Big(a_{j;d}(c) \cos(\pi j \widetilde{x}^{(d)}) + b_{j;d}(c) \sin(\pi j \widetilde{x}^{(d)}) \Big), \tag{6}$$

где J – число используемых элементов в ряде Фурье, а веса вычисляются на основе доступной обучающей выборки

$$a_{j;d}(c) = \frac{J+1-j}{(J+1)N(c)} \sum_{n=1}^{N(c)} \cos(\pi j \widetilde{x}_{c;n}^{(d)}), \tag{7}$$

$$b_{j;d}(c) = \frac{J+1-j}{(J+1)N(c)} \sum_{n=1}^{N(c)} \sin(\pi j \widetilde{x}_{c;n}^{(d)}). \tag{8}$$

Выигрыш в вычислительной сложности алгоритма классификации с проекционными оценками (1), (4), (6)-(8) по сравнению с ВНС (1), (3) происходит за счет того, что параметр J оказывается намного ниже числа $N(c)$ обучающих примеров c-го класса. Как известно [19], оценка вида (6) сходится со скоростью сходимости $O((N(c)^{-2/3})$, если $J = O(\sqrt[3]{N/C})$. В таком случае вычислительная сложность в худшем случае, когда на каждом этапе последовательного анализа сохраняются все C классов, составляет $O(N^{1/3}C^{2/3}D)$, что в $(N/C)^{2/3}$ раз ниже сложности традиционной ВНС [20]. В лучшем случае уже после первого ($l = 1$) шага в последовательном анализе получается единственное решение ($C_2 = \{c_1^*\}$), тогда сложность предлагаемого алгоритма окажется в L раз меньше и составит $O(N^{1/3}C^{2/3}D/L)$.

2.3. Досрочный останов прямого прохода в СНС. Рассмотренный в предыдущих разделах алгоритм может использоваться только для ускорения классификации при наличии вектора признаков $\mathbf{x}$, который извлекается с помощью прямого прохода (inference) по всей СНС. К сожалению, в таком случае время извлечения признаков $\overline{t}_{inf}$ может на порядок превышать время классификации $\overline{t}_{cls}$. Поэтому для повышения вычислительной эффективности в настоящем разделе рассмотрим реализацию возможности досрочного останова прохода в нейронной сети при распознавании изображений на основе последовательного анализа [17]. Для этого выберем $M > 1$ промежуточных слоев сети [21] так, чтобы они разбивали весь вычислительный граф СНС на M последовательно связанных частей:

$$\mathbf{z}_m = f_{exit_m}(\mathbf{z}_{m-1}), m \in \{1, \ldots, M\}, \tag{9}$$

где $\mathbf{z}_0$ – матрица пикселей входного изображения X, а f_{exit_m} – выход m-го промежуточного слоя. В связи с тем, что выход сверточного слоя является многомерным тензором, для его преобразования в вектор признаков $\mathbf{x}_m$ в настоящей работе предлагается добавить слой GAP (Global Average Pooling). В результате входное изображение описывается в виде иерархии признаков $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M$. Для каждого m-го выхода можно обучить отдельный классификатор, используя

векторы признаков, извлеченные на m-м слое из эталонного изображения X_n. Далее предположим, что каждый классификатор описывается C-мерным вектором $\mathbf{s}_m(\mathbf{x}_m) = [s_m^{(1)}(\mathbf{x}_m), \ldots, s_m^{(C)}(\mathbf{x}_m)]$ степеней уверенности (confidence score) $s_m^{(c)}(\mathbf{x}_m)$. Например, для рассмотренной в предыдущем разделе модификации ВНС для каждого класса c может использоваться оценка плотности распределения (4), (6), а для SVM (support vector machine) следует вычислять расстояние от опорных векторов класса до разделяющей гиперплоскости.

В рамках последовательного анализа прямой проход по участку СНС между m-м и $(m+1)$-м слоями (9) выполняется только в том случае, если не удалось получить достаточно надежное решение для вектора признаков $\mathbf{x}_m$. Досрочный останов происходит, если

$$\max_{c\in\{1,\ldots,C\}} s_m^{(c)}(\mathbf{x}_m) > s_m. \tag{10}$$

Для оценки порогов s_m воспользуемся леммой Неймана-Пирсона, в которой необходимо зафиксировать уровни значимости α_m на каждом m-м слое. В рамках теории множественных сравнений [1] выбирается уровень значимости α всей адаптивной процедуры принятия решений. Предполагая, что надежность классификации увеличивается при переходе к следующим слоям сети, можно воспользоваться процедурой Бенджамини-Хохберга [1]: $\alpha_m = \alpha \cdot m/M$. Далее наугад выбираются $0 < K < N$ примеров из обучающей выборки с номерами $\{n_1, \ldots, n_K\}$, которые используются для первоначального обучения m-го классификатора. Порог s_m оценивается на основе его предсказаний для множества оставшихся $(N-K)$ эталонов как α_m-квантиль максимальных степеней уверенности $\left\{ \max\limits_{c\neq c(n)} s_m^{(c)}(\mathbf{x}_{n;m}) \middle| n \notin \{n_1, \ldots, n_K\} \right\}$ [17].

Таким образом, алгоритм последовательного анализа в СНС (Рис. 1) заключается в следующем. На первом этапе входное изображение X подается на вход первой части нейронной сети (до $(m=1)$-го промежуточного слоя) и вычисляются тензор $\mathbf{z}_m$ (9) и соответствующий ему вектор признаков $\mathbf{x}_m$, подаваемый на вход m-го классификатора. Если максимальная степень уверенности превышает порог s_m (10), поиск останавливается. В противном случае $\mathbf{z}_m$ подается на вход следующего участка СНС между m-м и $(m+1)$-м слоями (9), и процедура последовательного анализа повторяется до тех пор, пока не получено достаточно надежное решение (10) или не достигнут последний слой

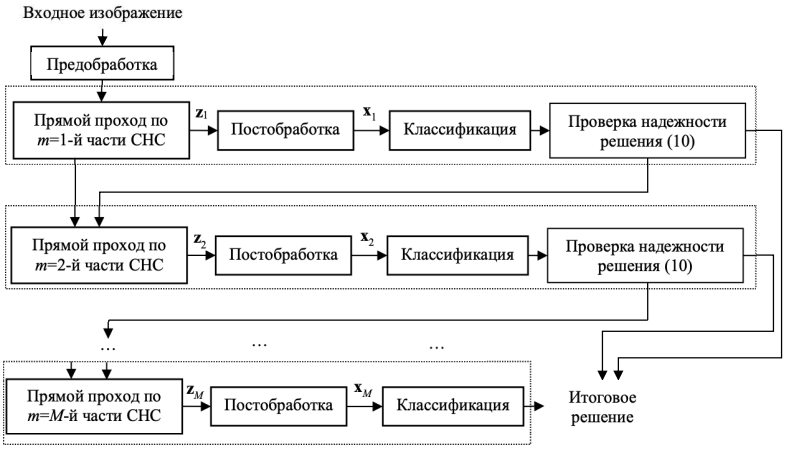

Рис. 1. Процесс последовательного анализа в СНС.

сети. В последнем случае можно либо отложить принятие итогового решения, либо вернуть максимально правдоподобный класс (1).

2.4. Последовательное дообучение СНС. Зачастую в задачах распознавания изображений объем доступной выборки не достаточен для обучения сложной нейросетевой модели. В таком случае наиболее часто используется перенос знаний (transfer learning) [6], в котором на первом шаге СНС обучается на основе сверхбольшой коллекции дополнительно собранных изображений. Для распознавания произвольных изображений обычно применяется предварительное обучение сети с помощью набора данных ImageNET. А для задачи обработки изображений лиц можно собрать большую коллекцию свободно доступных фотографий и предварительно решать задачу идентификации лиц [2]. В СНС выход из $D \gg 1$ значений предпоследнего слоя поступает на вход последнего полносвязного слоя, на котором и принимается решение в пользу одного из классов этой коллекции. Такую архитектуру можно рассматривать как применение логистической регрессии (ЛР) в последнем слое сети для классификации D признаков, выделенных на предыдущих слоях. Поэтому на втором шаге последний полносвязный слой заменяется на новый слой с C выходами (по одному на каждый

класс исходной задачи), и происходит дообучение (fine-tuning) полученного таким образом нейросетевого классификатора для доступного обучающегося множества из N эталонов [6, 13].

Стоит отметить, что на практике зачастую необходимо одновременно решать сразу несколько схожих задач распознавания изображений. Поэтому в настоящей статье предлагается выполнить последовательное обучение нескольких нейросетевых моделей так, чтобы на каждом очередном этапе выполнялось дообучение СНС, полученной на предыдущем этапе [7]. В качестве примера рассмотрим обучение эффективных по затратам памяти и времени принятия решений в таких моделях, как MobileNet, для задач анализа изображений лиц. На первом этапе набор данных VGGFace2 с 3 миллионов фотографий более чем 9000 людей [2] использована для обучения СНС в задаче идентификации лиц. Извлекаемые такой сетью дескрипторы могут использоваться в разнообразных задачах верификации, идентификации и кластеризации лиц [14].

На следующем шаге к этой модели добавляется полносвязный слой и два новых выхода для распознавания пола и возраста человека. При этом базовая часть сети для извлечения признаков фиксируется, и ее веса остаются неизменными. Для обучения использовались более 300 тысяч фотографий лиц из набора данных IMDB-Wiki [11]. К сожалению, возрастные группы в нем крайне не сбалансированы, что приводит к низкому качеству предсказания возраста для очень юных или пожилых людей. Поэтому к обучающему множеству были добавлены 15 тысяч фотографий из набора данных Adience [5]. Так как последний содержит только возрастные диапазоны, например, (0-2), (60-100), все изображения из таких диапазонов были сопоставлены со средним возрастом диапазона, в данном случае, 1 и 80. К сожалению, даже в таком случае дисбаланс классов зачастую приводит к неточным предсказаниям возраста. Поэтому для принятия окончательного решения предлагается выбрать top-K ($K \in \{1, 2, \ldots, C_a\}$) возрастов $\{a_1, \ldots, a_K\}$ с максимальными степенями уверенности $p_{a_k}(X)$ на выходе СНС, где C_a – число различных классов возраста ($C_a = 100$ для набора IMDB-Wiki). Далее только эти предсказания использованы для

вычисления математического ожидания возраста:

$$\overline{a}(X) = \frac{\sum\limits_{k=1}^{K} a_k \cdot p_{a_k}(X)}{\sum\limits_{k'=1}^{K} p_{a_{k'}}(X)}. \tag{11}$$

На следующем шаге последовательного обучения СНС рассматривается задача распознавания расы. В связи с тем, что, наряду с полом, раса является одним из постоянных характеристик в обучающих наборах данных, было принято решение расширить нейросетевую модель дополнительным выходом для классификации расы [3], при этом остальные веса сети в процессе обучения не менялись. Для настройки классификатора использовался набор данных University of Tennessee, Knoxville Face Dataset (UTKFace) [24], содержащий 5 классов (White, Black, Asian, Indian, Latino/Middle Eastern).

Наконец, на заключительном шаге была обучена СНС для классификации эмоций лиц на статических изображениях. К сожалению, в этом случае невозможно использовать дескриптор лиц, обученный для задачи идентификации, в качестве вектора признаков для высокоточной классификации эмоций. Действительно, при распознавании лиц разные эмоции одного и того же человека не должны оказывать существенного влияния на получаемый дескриптор. Поэтому в настоящей работе предлагается выполнить дообучение *всех* весов СНС с использованием достаточно большого набора эмоциональных лиц, например, AffectNet [8], содержащего 8 категорий (Anger, Contempt, Disgust, Fear, Happiness, Neutral, Sadness, Surprise). В связи с несбалансированностью классов использовалась взвешенная функция потерь NLL (negative log-likelihood), где вес каждой эмоции обратно пропорционален числу ее примеров в обучающем множестве. В результате такого последовательного обучения было получено две различные модели вида MobileNet v1: одна для одновременного распознавания пола, возраста, расы и извлечения характерных признаков, а также вторая модель для классификации эмоций. Несмотря на то, что вторая СНС была получена с помощью дообучения первой, обе могут использоваться независимо друг от друга. Для демонстрации примера их

применения было разработано демонстрационное мобильное Android-приложение[1].

§3. Результаты экспериментального исследования

3.1. Классификация изображений. Рассмотрим применение описанных выше методов последовательного анализа для распознавания изображений из набора данных Caltech-101 Object Category, включающего в себя 8677 изображений $C = 101$ классов. Для тестирования обучающее множество формировалось из 10 наугад выбранных примеров каждого класса, а тестовое множество состояло из всех остальных изображений. Для замеров времени принятия решений применялся сервер (12 core AMD Ryzen Threadripper, DDR4 64Gb, 3 NVIDIA GeForce GTX 1080Ti). Для извлечения признаков использовались предпоследние (GAP) слои предварительно обученных CHC Inception v3, InceptionResNet v2 и EfficientNet-b5 (Rand-aug). Во всех экспериментах применялась библиотека TensorFlow 2.0.

В первом эксперименте предложенный в разделе 2.1 алгоритм классификации на основе проекционных оценок (4), (6)–(8) и последовательного анализа (1)–(2) сопоставлялся с традиционными методами: k-ближайших соседей (k-БС с параметром $k = 3$), SVM (с RBF ядром) и базовый ВНС (1), (3). Для ускорения вычислений во всех случаях извлекались только $\widetilde{D} = 256$ главные компоненты. В предлагаемом последовательном анализе использовался порог $\delta = 0.9$, а на каждом шаге проверялось $m = 32$ компонент вектора признаков. Средняя точность $\overline{Acc}$ и время классификации $\overline{t}_{cls}$ (без учета времени извлечения признаков) представлены в Таблице 1.

Здесь, во-первых, стоит отметить высокую точность архитектуры EfficientNet, на 5–6% превосходящую остальные модели. Во-вторых, методы, основанные на полном переборе всех эталонов (k-БС, ВНС) ожидаемо оказались самыми медленными. Наконец, в-третьих, предлагаемый подход на основе последовательного анализа оказался весьма эффективным способ повышения (в 5–20 раз) скорости классификации, причем как в реализации на основе традиционной ВНС (3), так и для проекционных оценок плотности вероятности главной компоненты каждого класса (6)–(8).

[1] `https://github.com/HSE-asavchenko/MADE-mobile-image-processing/tree/master/lesson6/src/FacialProcessing`

Таблица 1. Результаты классификации изображений

Классификатор	Inception		InceptionResNet		EfficientNet	
	$\overline{Acc}$, %	$\overline{t}_{cls}$, мс	$\overline{Acc}$, %	$\overline{t}_{cls}$, мс	$\overline{Acc}$, %	$\overline{t}_{cls}$, мс
k-БС	84,56	1,02	84,13	0,96	89,55	3,98
SVM	82,97	0,25	82,52	0,25	88,20	0,99
ВНС	84,39	0,83	83,56	0,91	89,63	3,50
Последовательный анализ (1)-(2),(3)	83,92	0,16	83,54	0,14	89,05	0,63
Последовательный анализ (1)-(2),(6)	86,48	0,04	84,96	0,04	90,42	0,12

Таблица 2. Результаты досрочного останова прямого прохода

Классификатор	Слой	InceptionResNet v2		ResNet-152	
		$\overline{Acc}$, %	$\overline{t}_{inf}$, мс	$\overline{Acc}$, %	$\overline{t}_{inf}$, мс
ЛР	$m = 1$	10,75	25,71	18,86	9,28
ЛР	$m = M$	88,47	34,39	92,19	29,65
SVM	$m = 1$	92,82	25,82	55,49	8,91
SVM	$m = M$	95,04	34,51	93,09	38,44
BranchyNet [21]		87,48	32,22	79,10	33,67
CDL [9]		87,12	31,79	78,32	32,31
Последовательный анализ		95,19	27,25	92,67	30,10

В следующем эксперименте исследовалось применение последовательного анализа для ускорения прямого прохода по СНС (Раздел 2.3). В качестве промежуточных выходов наряду с последним слоем использовались слои block17_17_ac и block8_5_ac в InceptionResNet v2 и слой conv4_block1_out в ResNet-152. Было реализовано дообучение моделей с несколькими выходами (по одному на каждый промежуточный слой) следующим образом. В течение 5 эпох обучались только добавленные выходные слои, после чего в течение 10 эпох обучались все веса. Предложенный подход на основе последовательного анализа (9)–(10) с уровнем значимости $\alpha = 1\%$ сопоставлялся с обычными классификаторами векторов признаков на выходе каждого промежуточного слоя, а также с такими адаптивными нейронными сетями, как BranchyNet [21] и CDL (Conditional Deep Learning) [9]. Средняя точность $\overline{Acc}$ и время распознавания $\overline{t}_{inf}$ (с учетом времени прямого прохода) представлены в Таблице 2.

Как показал этот эксперимент, традиционные методы [9, 21] характеризуются достаточно низкой точностью по сравнению с классификацией признаков на предпоследнем слое сети. В то же время использованием последовательного анализа позволило значимо (на 7–8 мс) снизить время принятия решений без существенных потерь в точности классификации.

3.2. Распознавание атрибутов лиц. В начале исследовалась задача классификации пола и возраста для набора данных UTKFace (In the Wild) [24]. Лица на изображениях детектировались и выравнивались с помощью функции get_face_chip из библиотеки DLib (параметр margin=0.4). Предложенная модель одновременного распознавания пола и возраста multi-task MobileNet [14], предобученная для идентификации лиц из набора данных VGGFace2, сопоставлялась с двумя отдельными моделями (single task) предсказания пола и возраста, а также multi-task MobileNet, предобученной на ImageNet. Все эти СНС дообучались на одном и том же множестве изображений, полученном объединением наборов данных IMDB-Wiki и Adience. Кроме того, использовались следующие свободно доступные модели распознавания пола и возраста: VGG16 (Deep expectation, DEX) [11] для предсказания возраста и пола, MobileNet2 (Agegendernet)[2], FaceNet[3] [18] и ResNet-50 (InsightFace)[4] [4].

Оценки точности распознавания пола, MAE (mean average error) предсказания пола и суммарное время на распознавание пола и возраста приведены в Таблице 3. Как видно, предлагаемое предобучение модели для задачи идентификации лиц из большого набора данных VGGFace2 позволяет существенно повысить качество классификации по сравнению с известными нейросетевыми моделями. При этом вычислительная эффективность алгоритма обработки одного изображения лица оказывается очень высокой за счет получения двух решений за один проход по нейронной сети.

В следующем эксперименте исследовалась точность распознавания расы для набора UTKFace, который был разделен на 20149 обучающих и 3559 изображений. Последовательно обученная модель MobileNet сопоставлялась с традиционными классификаторами для известных дескрипторов лиц VGGFace (VGG-16) [10], VGGFace-2 (ResNet-50) [2] и

[2]https://github.com/dandynaufaldi/Agendernet

[3]https://github.com/BoyuanJiang/Age-Gender-Estimate-TF

[4]https://github.com/deepinsight/InsightFace/

Таблица 3. Результаты распознавания пола и возраста, набор данных UTKFace

Модель	Точность (пол), %	MAE (возраст)	$\overline{t}_{inf}$, мс
DEX	891,05	6,48	47,1
MobileNet2 (Agegendernet)	91,47	7,29	11,4
FaceNet	89,54	8,58	20,3
ResNet-50 (InsightFace)	87,52	8,57	25,3
Multi-task MobileNet (ImageNet)	91,81	5,88	4,7
Single-task MobileNet (VGGFace2)	93,59	5,94	7,2
Multi-task MobileNet (VGGFace2)	94,10	5,44	4,7

Таблица 4. Точность (%) распознавания расы, набор данных UTKFace

Классификатор	VGGFace	VGGFace-2	FaceNet	**Multi-task MobileNet**
Random Forest	83,5	87,8	84,3	83,8
k-NN	76,2	84,5	84,4	82,2
SVM (RBF)	78,8	82,4	86,2	87,7
SVM (linear)	79,5	83,1	85,6	85,6
Полносвязный слой	80,4	86,4	84,4	87,0

FaceNet [18]. В Таблице 4 приведены оценки точности каждого классификатора. Как видно из этой таблицы, последовательно обученная модель оказалась достаточно точной по сравнению с архитектурами с более высокими затратами памяти и вычислительной сложностью. При этом, как известно [14], такие модели для исходных задач распознавания лиц оказываются намного более точными по сравнению с легковесной MobileNet.

В заключительном эксперименте рассмотрим результаты классификации эмоций для тестового набора данных AffectNet. Для настройки всех моделей использовалось обучающее множество AffectNet. Предлагаемая последовательно дообученная MobileNet сопоставлялась как с традиционными СНС (MobileNet v1, Inception v3, EfficientNet B3), предобученными на ImageNet-1000, так и с дескрипторами лиц VGGFace

Таблица 5. Точность распознавания эмоций, набор данных AffectNet

СНС	Точность, %
Дообученная MobileNet	56,88
Дообученная Inception v3	59,65
Дообученная EfficientNet B3	60,28
Дообученная VGG-16 (VGGFace)	54,78
Предобученная ResNet-50 (VGGFace2)	40,87
Дообученная ResNet-50 (VGGFace2)	57,01
Предобученная Multi-task MobileNet (VGGFace2)	29,13
Дообученная Multi-task MobileNet (VGGFace2)	60,43

и VGGFace2. Последние рассматривались в двух вариантах: простая классификация исходных дескрипторов (веса базовой сети не дообучались) и дообучением всей сети целиком. Основные результаты этого эксперимента приведены в Таблице 5. Как видно, предлагаемый подход с последовательным дообучением позволил добиться наивысшей точности на тестовом множестве, которая на 4,5% превосходит точность традиционного дообучения MobileNet.

§4. Заключение

В настоящей статье рассмотрено применение последовательного анализа для повышения эффективности распознавания изображений. Показано (Таблица 1), что последовательный анализ главных компонент векторов признаков, извлеченных с помощью СНС, оказывается в 5-20 раз быстрее базовых методов instance-based learning (k-NN/ВНС). В то же время применение последовательного анализа для досрочного останова прямого прохода по нейронной сети, хоть и позволяет иногда повысить скорость принятия решений (Таблица 2), оказывается недостаточно универсальным. Например, его применение совместно с моделями MobileNet для классификации атрибутов лиц не привело к значимому повышению быстродействия в связи с малой точностью решений, полученных при классификации векторов признаков, которые извлечены из ранних слоев СНС. Наконец, экспериментально продемонстрировано, что последовательный подход к дообучению нейронных сетей для сходных задач распознавания пола, возраста, расы и эмоций по фотографии лица приводит к получению

высокоточных легковесных моделей MobileNet, значительно превосходящих по вычислительной эффективности более глубокие архитектуры ResNet/Inception/EfficientNet (Таблицы 3, 4, 5).

Список литературы

1. Y. Benjamini, Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* — J. the Royal statistical society: series B (Methodological) **57** (1995), No. 1, 289–300.
2. Q. Cao, Li Shen, W. Xie, O. M. Parkhi, A. Zisserman, *VGGFace2: A dataset for recognising faces across pose and age*, in: Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 67–74, 2018.
3. A. Das, A. Dantcheva, F. Bremond, *Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach*, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, pp. 573–585, 2018.
4. J. Deng, J. Guo, X. Niannan, S. Zafeiriou, *Arcface: Additive angular margin loss for deep face recognition*, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4690-4699, 2019.
5. E. Eidinger, R. Enbar, T. Hassner, *Age and gender estimation of unfiltered faces.* — IEEE Transactions on Information Forensics and Security **9** (2014), No. 12, 2170–2179.
6. I. Goodfellow, Y. Bengio, A. Courville, *Deep learning*, MIT press, 2016.
7. S. Hung, J.-H. Lee, T. Wan, C.-H. Chen, Y.-M. Chan, C.-S. Chen, *Increasingly packing multiple facial-informatics modules in a unified deep-learning model via lifelong learning*, in: Proceedings of the 2019 on International Conference on Multimedia Retrieval (ICMR), pp. 339–343, 2019.
8. A. Mollahosseini, B. Hasani, M. H. Mahoor, *AffectNet: A database for facial expression, valence, and arousal computing in the wild.* — IEEE Transactions on Affective Computing **10** (2017), No. 1, 18–31.
9. P. Panda, A. Sengupta, K. Roy, *Conditional deep learning for energy-efficient and enhanced pattern recognition*, in: Proceedings of IEEE Design, Automation & Test in Europe Conference & Exhibition, pp. 475–480, 2016.
10. O. M. Parkhi, A. Vedaldi, A. Zisserman, *Deep face recognition*, in: Proceedings of the British Machine Vision Conference (BMVC), No. 3, 2015.
11. R. Rothe, R. Timofte, L. Van Gool, *DEX: Deep expectation of apparent age from a single image*, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 10–15, 2015.
12. A. V. Savchenko, *Search techniques in intelligent classification systems*, Springer, 2016.

13. А. В. Савченко, *Метод максимально правдоподобных рассогласований в задаче распознавания изображений на основе глубоких нейронных сетей*. — Компьютерная оптика **41** (2017), No. 3, 422–430.
14. A. V. Savchenko, *Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output ConvNet*. — PeerJ Computer Science **5** (2019), e197.
15. A. V. Savchenko, *Sequential three-way decisions in multi-category image recognition with deep features based on distance factor*. — Information Sciences **489** (2019), 18–36.
16. A. V. Savchenko, *Probabilistic neural network with complex exponential activation functions in image recognition*. — IEEE Transactions on Neural Networks and Learning Systems **31** (2020), No. 2, 651–660.
17. A. V. Savchenko, *Sequential analysis with specified confidence level and adaptive convolutional neural networks in image recognition*, in: Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN), 2020.
18. F. Schroff, D. Kalenichenko, J. Philbin, *FaceNet: A unified embedding for face recognition and clustering*, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823, 2015.
19. S. C. Schwartz, *Estimation of probability density by an orthogonal series*. — The Annals of Mathematical Statistics (1967), 1261–1265.
20. D. F. Specht, *Probabilistic neural networks*. — Neural Networks **3** (1990), No. 1, 109–118.
21. S. Teerapittayanon, B. McDanel, H.T. Kung, *BranchyNet: Fast inference via early exiting from deep neural networks*, in: Proceedings of the 23rd IEEE International Conference on Pattern Recognition (ICPR), pp. 2464–2469, 2016.
22. A. Wald, *Sequential Analysis*, Dover Publications, New York, 2013.
23. Y. Y. Yao, X. F. Deng, *Sequential three-way decisions with probabilistic rough sets*, in: Proceedings of ICCI*CC, IEEE Computer Society, pp. 120–125, 2011.
24. Z. Zhang, Y. Song, H. Qi, *Age progression/regression by conditional adversarial autoencoder*, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5810–5818, 2017.

Savchenko A. V. Fast image classification algorithms based on sequential analysis.

In this paper fast image recognition techniques based on statistical sequential analysis are discussed. We examine the possibility to sequentially process the principal components and organize a convolutional neural network with early exits. Particular attention is paid to sequentially learn multi-task lightweight neural network model to predict several facial attributes (age, gender and ethnicity) based on preliminary training on the face classification task. It is highlighted that the whole above-mentioned model should be fine-tuned in order to deal with emotion recognition problem. Experimental study on several datasets demonstrate that the

proposed approach is rather accurate and has very low run-time and space complexity when compared to known state-of-the-art methods.

Национальный
исследовательский университет
Высшая школа экономики,
Нижний Новгород,
Россия, 603155

E-mail: avsavchenko@hse.ru

Поступило 20 августа 2020 г.