© 2022 г. Д.В. СВИТОВ (d.svitov@expasoft.tech)
(ООО Экспасофт, Новосибирск;
Институт автоматики и электрометрии СО РАН, Новосибирск),
С.А. АЛЯМКИН, канд. техн. наук (s.alyamkin@expasoft.com)
(ООО Экспасофт, Новосибирск)

ДИСТИЛЛЯЦИЯ МОДЕЛЕЙ ДЛЯ РАСПОЗНАВАНИЯ ЛИЦ, ОБУЧЕННЫХ С ПРИМЕНЕНИЕМ ФУНКЦИИ СОФТМАКС С ОТСТУПАМИ

Использование сверточных нейронных сетей в сочетании с функцией Софтмакс (анг. softmax) с отступами позволяет достичь наибольшей точности в задаче распознавания лиц. Развитие встраиваемых систем, таких как умные домофоны, породило интерес к легковесным нейронным сетям. Так были предложены облегченные нейросетевые модели, обученные с применением функции Софтмакс с отступами, для задачи идентификации по лицу. В данной работе предлагается метод дистилляции, который позволяет получить большую точность, чем другие методы для задачи распознавания лиц на наборах данных LFW, AgeDB-30 и Megaface. Основная идея предлагаемого подхода заключается в использовании центров классов сети-учителя для инициализации сети-ученика. Затем сеть-ученик обучается производить биометрические вектора, углы от которых до центров классов равны углам в сети-учителе.

Ключевые слова: сверточные нейронные сети, дистилляция, биоидентификация.

DOI: 10.31857/S000523102210004X, **EDN:** AJXJDG

1. Введение

В недавнее время большой интерес получила разработка систем распознавания лиц, требующих малых вычислительных ресурсов. Это вызвано широким распространением встраиваемых систем, таких как домофоны с функцией доступа по лицу и камеры наружного наблюдения. Такие системы распознавания лиц основываются на нейросетевых моделях для мобильных вычислителей. Во время работы нейросетевая модель получает на вход изображение лица и восстанавливает по нему вектор фиксированной длины. В таком векторе закодирована информация о лице на изображении, и он называется биометрическим. Чем дальше вектора для различных людей друг от друга в векторном пространстве и ближе для различных изображений одного человека, тем выше качество работы нейронной сети. Близость векторов определяется согласно выбранной метрике, в рассматриваемых в данной работе моделях это косинусное расстояние.

К таким моделям относится архитектура MobileFaceNet [1], разработанная специально для распознавания лиц на устройствах с малой вычислительной

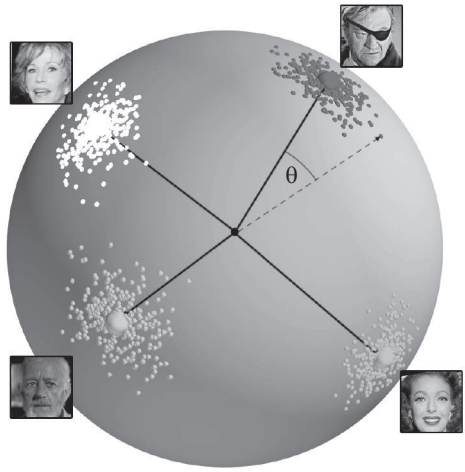

Рис. 1. Гиперсфера нормированных биометрических векторов. Различными цветами обозначены разные классы из обучающего набора данных: большими кругами обозначены центры классов, меньшими — изображения, относящиеся к классу. Для вычисления близости изображения к классу необходимо вычислить угол θ между векторами.

мощностью. В свою очередь использование функции Софтмакс с отступами [3–5] для обучения таких моделей позволяет достичь наибольшую точность в задачах идентификации и верификации по лицу. Нейросетевые модели для биоидентификации обучаются на классификацию с функцией Софтмакс большого количества классов, затем вектор, получаемый на предпоследнем слое, используется для кодирования и сравнения новых изображений. Добавление отступов в функцию Софтмакс в процессе обучения сети добавляет дополнительный отступ для векторов, относящихся к одному классу, что вынуждает сеть минимизировать угол от вектора до центра класса. Таким образом, сеть обучается минимизировать косинусное расстояние между изображениями одного человека (рис. 1).

Быстрые и компактные мобильные нейросетевые архитектуры достигают меньшей точности, чем серверные решения. В задачах биометрического доступа такое снижение точности может играть критическую роль. Для увеличения точности мобильных нейросетевых архитектур используется дистилляция [9]. Дистилляция нейронной сети — это метод передачи знаний от сети-учителя с большим числом обучаемых параметров в сеть-ученика с малым числом обучаемых параметров. В данной работе предлагается новый подход дистилляции, позволяющий сократить разницу в точности между сетью-учителем и сетью-учеником.

Идея предложенного подхода заключается в копировании последнего слоя сети, содержащего обученные центры классов обучающего набора данных, из сети-учителя в сеть-ученика и заморозке этого слоя во время всей процедуры дистилляции. Дистилляция заключается в обучении сети-ученика воспроизводить углы между центрами классов и биометрическими векторами для лиц, равные углам между соответствующими векторами и центрами классов

в сети-учителе. Такой подход позволяет сети-ученику лучше воспроизводить результат работы сети-учителя.

Основной вклад данной работы заключается в следующем:

1) предлагается новый метод для дистилляции нейронных сетей, обученных с применением функции Софтмакс с отступами;

2) предложенный метод позволяет сократить разницу в точности между сетью-учителем и сетью-учеником. Использование предложенного метода позволило для мобильной нейросетевой архитектуры получить наибольшую точность в сравнении с другими методами дистилляции для наборов данных LFW [6], AgeDB-30 [7] и MegaFace [8];

3) в данной работе производится прямое сравнение различных методов дистилляции. Программный код с реализацией рассматриваемых методов и проводимых экспериментов доступен в открытом доступе на сайте github.com [27].

2. Обзор литературы

2.1. Функция Софтмакс с отступами

Существует несколько вариаций функции Софтмакс (анг. softmax) с отступами, используемых при обучении нейронных сетей для систем распознавания лиц. Они включают подходы Cosface [5], Sphereface [4] и Arcface [3]. Все три подхода могут быть описаны общей формулой, задающей функцию ошибки классификации изображений лиц:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}m_1+m_2)-m_3)}}{e^{s(\cos(\theta_{y_i}m_1+m_2)-m_3)}+\sum\limits_{j=1,j\neq y_i}^{n}e^{s\,\cos\theta_j}}. \tag{1}$$

Данные подходы могут быть получены из представленной формулы (1) подстановкой значений параметров m_1, m_2, m_3. Подход Sphereface получается при $m_1 = 4$, $m_2 = m_3 = 0$; подход Cosface получается при $m_1 = 1$; $m_2 = 0$; $m_3 = 0{,}35$; подход Arcface получается при $m_1 = 1$; $m_2 = 0{,}5$; $m_3 = 0$. В представленной формуле N — это количество изображений, используемых на каждом шаге стохастического градиентного спуска; θ_j — угол между векторами, соответствующими обучающему примеру с индексом i и центру класса с индексом j; y_i — соответствует индексу класса, к которому относится пример с индексом i согласно разметке; s — константа масштабирования, во всех случаях имеющая значение 64. Детальный анализ представленной формулы можно найти в статье Arcface [3]. Из рассмотренных подходов Arcface демонстрирует наибольшую точность на наборах данных LFW, AgeDB-30 и MegaFace.

2.2. Дистилляция

Дистилляция знаний от сети-учителя к сети-ученику была впервые предложена в [9]. Данный подход заключается в обучении сети-ученика с малым числом обучаемых параметров за счет передачи знаний от сети-учителя с большим числом обучаемых параметров. Под передачей знаний понимается

передача некоторой информации от обученной более точной модели для улучшения сходимости обучаемой малой модели. Ключевая идея подхода, предложенного в [9], заключается в передаче знаний через приближение сглаженного вероятностного распределения на выходе сети-учителя. Это достигалось за счет добавления делителя в формулу функции Софтмакс.

Некоторые исследования продолжают развивать идею использования сглаженного распределения вероятности в качестве разметки для обучения сети-ученика. Так, например, в [10] был предложен подход к дистилляции ансамбля нейронных сетей в одну сеть-ученика, используя данный метод. В [11] предложен подход к обучению сети-ученика с помощью сети-учителя, подверженной зашумленности выхода. В [12] сеть-ученик и сеть-учитель обучались с одинаковой параметризацией. В [25] предлагается механизм дистилляции через задание априорного распределения сети-ученика на основе апостериорного распределения сети-учителя с измененной структурой модели для совпадения пространства параметров. Также делаются шаги в сторону теоретического обоснования методов дистилляции через вероятностную интерпретацию [25, 26].

Другой подход к дистилляции — это дистилляция через скрытые слои нейронной сети. В [13] предлагается обучать сеть-ученика копировать распределение весов на скрытых слоях сети-учителя. В [14] промежуточные слои сетей ученика и учителя используются для регуляризации обучения. В [15] для регуляризации во время дистилляции используется ограничение на сохранение взаимоотношения между локальными объектами. Для этого L_2 расстояние между биометрическими векторами для сети-ученика минимизируется на основе информации от сети-учителя. В [16] предлагается относительная дистилляция знаний, которая штрафует за структурные различия во взаимоотношении между обучающими примерами.

Для дистилляции легковесных нейронных сетей для задачи распознавания лиц, обученных с применением функции Софтмакс с отступами, применяются следующие подходы: триплетная дистилляция [17], дистилляция по углу [18] и дистилляция на основе отступа [19].

В *триплетной дистилляции* сеть-ученик обучается с триплетной функцией ошибки, отступ в которой вычисляется основываясь на расстоянии между якорным и негативным примерами и якорным и позитивным примерами, предсказанными сетью-учителем.

В *дистилляции по углу* минимизируется угол между биометрическими векторами сетей ученика и учителя для каждого примера в обучающей выборке.

В *дистилляции на основе отступа* предлагается производить дистилляцию через сглаженное распределение вероятности. Для этого в формулу (1) добавляется деление на параметр T по аналогии с [9].

3. Предлагаемый подход

3.1. Описание сети-учителя и сети-ученика

В качестве сети-учителя рассматривалась нейронная сеть с архитектурой ResNet100 [20]. Выбор данной архитектуры был обусловлен тем,

Таблица 1. Сравнение параметров рассматриваемых нейросетевых архитектур. Время работы сетей замерялось для входного изображения размерностью $112 \times 112 \times 3$ на процессоре Intel Xeon(R) CPU E3-1270 v3 @ 3,50GHz×8

	ResNet100	MobileFaceNet
Число операций с плавающей точкой / 10^9	24,2	0,44
Размер / Мегабайт	261,2	5,3
Число обучаемых параметров / 10^6	52,56	1,19
Время работы / Миллисекунд	$401 \pm 25,7$	$42,2 \pm 5,48$

что она содержит большое число обучаемых параметров и, следовательно, позволяет достигать высокой точности в задаче распознавания лиц. В качестве сети-ученика была выбрана недавно предложенная архитектура MobileFaceNet(ReLU) [1], содержащая менее 1 млн параметров и разработанная специально для решения задачи распознавания лиц на мобильных процессорах. Данная архитектура состоит из блоков предложенных в MobileNetV2 [2], но позволяет обрабатывать изображения вдвое быстрее за счет уменьшения пространственной размерности изображения на ранних слоях и использования меньшего числа фильтров на промежуточных слоях блоков. В описываемых экспериментах была сделана следующая модификация данной архитектуры: размерность выходного биометрического вектора была увеличена с 256 до 512 элементов для совпадения с размерностью архитектуры ResNet100. В табл. 1 приводится сравнение рассматриваемых архитектур.

3.2. Описание метода

Обозначим биометрический вектор сети-ученика для изображения с индексом i как $x_{S_i} \in R^D$, где D — размерность вектора. И обозначим через $x_{T_i} \in R^D$ биометрический вектор соответствующего изображения для сети-учителя. Матрицы весов последнего слоя для сетей ученика и учителя будем обозначать соответственно $W_S \in R^{D \times n}$ и $W_T \in R^{D \times n}$, где n — количество классов в обучающем наборе данных. Столбец матрицы с индексом j, соответствующий центру класса y_i, к которому относится изображение с индексом i из обучающего набора данных, обозначается как $W_{S_j} \in R^D$ для сети-ученика и $W_{T_j} \in R^D$ для сети-учителя.

Подходы, основанные на добавлении отступа m в функцию Софтмакс, производят нормировку столбцов матрицы весов и биометрических векторов на 1: $||W_j|| = 1$ и $||x_i|| = 1$, где W_j — это j-й столбец матрицы W, а x_i — это биометрический вектор, соответствующий i-му обучающему примеру. Такая нормировка позволяет рассматривать результаты произведений векторов на последнем слое сети как косинусные расстояния $\cos(\theta_j)$ между биометрическими векторами и соответствующими центрами классов: ${W_j}^{\mathrm{T}} x_i = ||W_j|| \times$ $\times ||x_i|| \cos(\theta_j) = \cos(\theta_j)$ (рис. 1), где θ_j — угол между векторами, соответствующими обучающему примеру с индексом i и центру класса с индексом j. Далее в качестве частного случая обучения с использованием функции Софтмакс с отступами рассматривается метод Arcface [3], задаваемый формулой,

подробно описанной в разделе “Обзор литературы”:

$$L_{\text{ArcFace}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))}+\sum\limits_{j=1,j\neq y_i}^{n}e^{s\,\cos\theta_j}}. \tag{2}$$

Данный метод был выбран для рассмотрения, так как позволяет получать наибольшую точность среди методов обучения с применением функции Софтмакс с отступами. В подходе Arcface значение отступа m фиксируется равным 0,5, как предложено в [3]. В данной работе предлагается производить дистилляцию знаний от сети-учителя через вычисление значения отступа m для каждого изображения i. Предлагаемый метод дистилляции содержит две основные идеи:

— Центры классов, найденные сетью-учителем, используются в сети-ученике: $W_S = W_T$. Так как центры классов обучаемые параметры, более глубокая сеть способна сформировать более хорошее их положение на гиперсфере. Такое положение, при котором относящиеся к этим центрам кластеры векторов будут разделимы с большей точностью.

— Для дистилляции знаний используются вычисляемые значения m_i для каждого изображения. Они явным образом контролируют расстояние между биометрическими векторами x_{S_i} и соответствующими центрами классов W_{S_j}. Большее значение m_i способствует приближению вектора x_{S_i} к центру класса.

Отступ m_i вычисляется, основываясь на угле между биометрическим вектором сети-учителя $x_{T_i} \in R^D$ и соответствующим вектором $W_{T_j} \in R^D$ центра класса, задаваемого как y_i. Отступ m для изображения с индексом i вычисляется аналогично отступу в триплетной дистилляции:

$$m_i = f(a_i) = \frac{m_{\max} - m_{\min}}{a_{\max}}a_i + m_{\min}, \tag{3}$$

$$a_i = \frac{W_{T_j}^{\mathrm{T}} x_{T_i}}{||W_{T_j}|| \cdot ||x_{T_i}||}, \tag{4}$$

где параметры $m_{\max} = 0{,}5$ и $m_{\min} = 0{,}2$ задают максимальное и минимальное значение отступа, по аналогии с формулой триплетной дистилляции, предложенной в [17]. А значение параметра $a_{\max}$ вычисляется как максимальное значение угла a в мини-пакете изображений на каждом шаге обучения.

Формула (4) используется для вычисления углов a_i между центрами классов W_{T_j} и биометрическими векторами x_{T_i} сети-учителя (рис. 2,*а*). Углы a_i, получаемые от сети-учителя, используются для вычисления отступа $m_i = f(a_i)$ по формуле (3). Чем меньше значение угла между биометрическим вектором и центром класса, к которому он относится, тем больший отступ m_i используется для этого вектора при обучении сети-ученика (рис. 2,*б*). Большее значение отступа вынуждает сеть минимизировать угол от биометрического вектора до центра класса, чтобы компенсировать эффект от отступа. Тем самым векторы, которые были близки к центру класса в сети-учителе, будут близки к центру класса в сети-ученике.

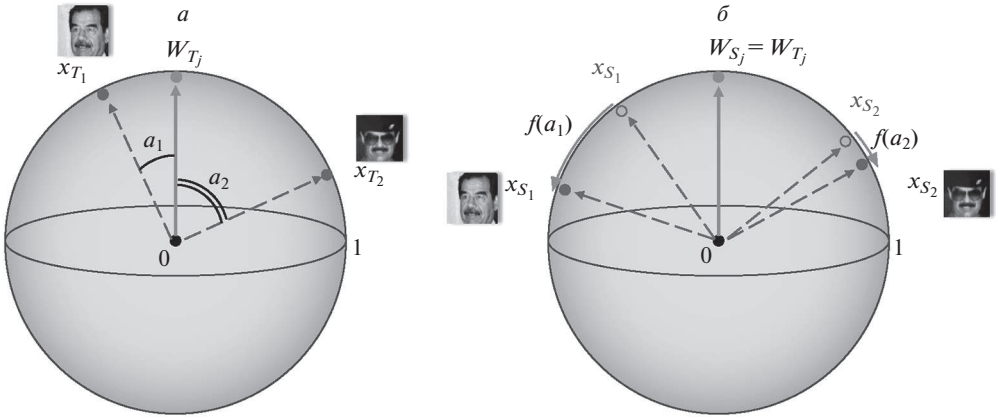

Рис. 2. *а* — Гиперсфера векторов сети-учителя. *б* — Вычисление сдвига для векторов на гиперсфере сети-ученика на основе углов до центра класса в сети-учителе.

Интуиция, лежащая в основе предлагаемого подхода, заключается в том, чтобы изменять биометрический вектор сети-ученика для уменьшения расстояния до центра класса, если соответствующий вектора сети-учителя находится близко к центру класса. Это позволяет осуществлять передачу знаний от сети-учителя к сети-ученику более эффективно потому, что сеть-ученик концентрируется на более уверенно классифицированных сетью-учителем изображениях.

Предлагаемый подход к дистилляции позволяет передавать информацию об относительном положении векторов на гиперсфере, не вводя явных ограничений на биометрические векторы сети-ученика.

4. Эксперименты

4.1. Детали реализации

Предобработка данных. Для обнаружения и выравнивания лиц на изображениях применялся метод MTCNN [21]. В качестве обучающего набора данных использовался набор данных для классификации по лицу MS1MV2 [3]. Данный набор данных — это очищенный в полуавтоматическом режиме набор данных MS-Celeb-1M [22], предложенный в работе, посвященной Arcface. Рассматриваемый набор данных содержит 5,8 млн изображений для 85 тыс. классов.

После процедуры выравнивания лиц, выполненной с применением ключевых точек найденных MTCNN, получаются изображения размером 112×112 пикселей. Значения яркости пикселей полученных изображений затем нормируются до диапазона $[-1, 1]$.

Процедура обучения. В качестве сети-учителя была обучена нейронная сеть с архитектурой ResNet100 и функцией Arcface. Сравнение всех методов дистилляции проводилось в одинаковых условиях, где передача знаний осуществлялась от ResNet100 к MobileFaceNet(ReLU). Дистилляция производи-

лась со следующими значениями гиперпараметров: размер мини-пакета изображений равнялся 512, шаг обучения равнялся 0,1 и увеличивался в 10 раз после 10 000, 160 000 и 200 000 итераций. Оптимизация обучаемых параметров выполнялась алгоритмом SGD [24] со значением моментов, равным 0,9. Коэффициент регуляризации обучаемых параметров равнялся $5e-4$. Значения максимального и минимального возможных значений отступа фиксировались равными $m_{\max} = 0{,}5$ и $m_{\min} = 0{,}2$. Значение параметра s метода Arcface не изменялось и равнялось 64. Описанные далее эксперименты основываются на официальной реализации Arcface авторами на фреймворке MXNet.

Для сравнения предложенного метода дистилляции с существующими подходами были реализованы триплетная дистилляция [17], дистилляция по углу [18] и дистилляция на основе отступа [19] на фреймворке MXNet. Нейронные сети дистиллировались с помощью этих методов с указанными в опорных статьях параметрами. Реализация этих методов на MXNet доступна в репозитории данной работы на GitHub [27].

Процедура тестирования. Обученные с помощью дистилляции нейронные сети для распознавания лиц тестировались на задачах верификации и идентификации.

Верификация. Для замеров точности обученных нейронных сетей на задаче верификации использовались наборы данных LFW и AgeDB-30. Каждый набор данных содержит порядка 3000 позитивных и 3000 негативных пар изображений. В процедуре тестирования обученная нейронная сеть использовалась для получения биометрических векторов для пары изображений лиц. Верификация производилась основываясь на косинусном расстоянии между векторами. Точность измерялась как процент верно верифицированных пар изображений.

Идентификация. Наиболее представительным и сложным протоколом тестирования для задачи идентификации лиц является MegaFace. В набор данных MegaFace входит миллион изображений лиц для 690 000 человек для формирования негативных примеров. И 100 000 изображений для 530 людей из набора данных FaceScrub [23], идентификацию которых по базе лиц необходимо произвести. Значением метрики качества является топ-1 точность для задачи идентификации по базе лиц с миллионом негативных примеров.

4.2. Результаты тестирования

Как показано в табл. 2, обученная с функцией Arcface, сеть-учитель достигает точности 99,76% для набора данных LFW и 98,21% для AgeDB-30. сеть-ученик, обученная с функцией Arcface, достигает 99,51% для набора данных LFW и 96,13% для AgeDB-30. Предложенный подход позволяет сократить разницу в точности сильнее, чем остальные рассмотренные подходы: до 99,61% для LFW и 96,55% для AgeDB-30.

В табл. 3 представлены результаты замеров точности для задачи идентификации по протоколу MegaFace с миллионом негативных примеров в базе лиц. Для данного протокола тестирования сеть-учитель достигает точности 98,35%, а сеть-ученик 90,62%. Предложенный подход позволяет увеличить точность идентификации до 91,70%. Также для задачи идентификации бы-

Таблица 2. Точность верификации на наборах данных LFW и AgeDB-30. В экспериментах использовалась версия MobileFaceNet с функцией активации ReLU

Архитектура	Метод обучения	LFW %	AgeDB-30 %
ResNet100 (сеть-учителя)	ArcFace [3]	99,76	98,21
MobileFaceNet (сеть-ученик)	ArcFace [3]	99,51	96,13
MobileFaceNet	Триплетная дистилляция по L2 [17]	99,56	96,23
MobileFaceNet	Триплетная дистилляция по cos [17]	99,55	95,60
MobileFaceNet	Дистилляция с отступом для T=4 [19]	99,41	96,01
MobileFaceNet	Дистилляция по углу [18]	99,55	96,01
MobileFaceNet	Предлагаемый метод	**99,61**	**96,55**

Таблица 3. Точность идентификации с использованием протокола MegaFace с 1 млн негативных примеров. В экспериментах использовалась версия MobileFaceNet с функцией активации ReLU

Архитектура	Метод-обучения	MegaFace %
ResNet100 (сеть-учитель)	ArcFace [3]	98,35
MobileFaceNet (сеть-ученик)	ArcFace [3]	90,62
MobileFaceNet	Триплетная дистилляция по L2 [17]	89,10
MobileFaceNet	Триплетная дистилляция по cos [17]	86,52
MobileFaceNet	Дистилляция с отступом для T=4 [19]	90,77
MobileFaceNet	Дистилляция по углу [18]	90,73
MobileFaceNet	Предлагаемый метод	**91,70**

ло замечено уменьшение точности для метода триплетной дистилляция, но данный метод показал хорошую точность для задачи верификации.

5. Оценка метода

Для более детального анализа предлагаемого подхода была проведена его оценка методом удаления различных элементов. При удалении элементов подхода оценивалось их влияние на точность верификации на наборе данных LFW. В табл. 4 оценивается влияние следующих аспектов метода:

— Копирование центров — копирование центров классов, представленных обучаемыми параметрами последнего слоя нейронной сети, от сети-учителя к сети-ученику: $W_S = W_T$.

— Использование m_i вместо m-использование вычисляемых m_i для дистилляции через Arcface вместо фиксированного $m = 0{,}5$.

— Фиксирование центров — пометка центров классов сети-ученика W_S как необучаемые параметры. Чтобы в процессе дистилляции они оставались равными центрам классов сети-учителя W_T.

Наибольший прирост в точности в 0,9% вызван копированием центров классов сети-учителя в сеть-ученика и дальнейшая пометка их как необуча-

Таблица 4. Оценка точности метода на наборе данных LFW удалением различных его элементов

Копирование центров	Использование m_i вместо m	Фиксирование центров	LFW %
√	√	√	**99,61**
√	√		98,31
√			99,55
	√		99,43
√		√	99,60

емые параметры. Самостоятельное обучение сетью-учеником центров классов W_S во всех сценариях ведет к уменьшению точности на наборе данных LFW. Соответственно ключевую роль в предложенном методе дистилляции играет копирование центров классов. Использование адаптивного отступа m_i позволяет дополнительно увеличить точность получаемой модели.

6. Заключение

В данной работе был предложен подход к дистилляции нейронных сетей, обученных для задачи распознавания лиц с функцией Софтмакс с отступами. Была продемонстрирована эффективность использования центров классов сети-учителя в сети-ученике. Было проведено сравнение предложенного метода с другими методами дистилляции для нейронных сетей, использующих Софтмакс с отступами. Описанный в данной работе подход позволяет получить лучшую точность на наборах данных LFW и AgeDB-30 для задачи верификации и для MegaFace для задачи идентификации.

Предложенный метод дистилляции может применяться для увеличения точности нейросетевых моделей с малым числом параметров для встраиваемых устройств. Таких, например, как камеры наружного наблюдения или домофоны с функцией доступа по лицу.

СПИСОК ЛИТЕРАТУРЫ

1. *Chen S., Liu Y., Gao X., Han Z.* Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices // Chinese Conference on Biometric Recognition. Springer, Cham, 2018. С. 428–438.
2. *Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L.C.* Mobilenetv2: Inverted residuals and linear bottlenecks // Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. С. 4510–4520.
3. *Deng J., Guo J., Xue N., Zafeiriou S.* Arcface: Additive angular margin loss for deep face recognition // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. С. 4690–4699.
4. *Liu W., Wen Y., Yu Z., Li M., Raj B., Song L.* Sphereface: Deep hypersphere embedding for face recognition // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. С. 212–220.

5. *Wang H., Wang Y., Zhou Z., Ji X., Gong D., Zhou J., Li Z., Liu W.* Cosface: Large margin cosine loss for deep face recognition // Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. С. 5265–5274.
6. *Huang G.B., Mattar M., Berg T., Learned-Miller E.* Labeled faces in the wild: A database forstudying face recognition in unconstrained environments // Workshop on faces in'Real-Life'Images: detection, alignment, and recognition. 2008.
7. *Moschoglou S., Papaioannou A., Sagonas C., Deng J., Kotsia I., Zafeiriou S.* Agedb: the first manually collected, in-the-wild age database // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017. С. 51–59.
8. *Kemelmacher-Shlizerman I., Seitz S.M., Miller D., Brossard E.* The megaface benchmark: 1 million faces for recognition at scale // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. С. 4873–4882.
9. *Hinton G., Vinyals O., Dean J.* Distilling the knowledge in a neural network // arXiv preprint arXiv:1503.02531. 2015.
10. *Fukuda T., Suzuki M., Kurata G., Thomas S., Cui J., Ramabhadran B.* Efficient Knowledge Distillation from an Ensemble of Teachers // Interspeech. 2017. С. 3697–3701.
11. *Sau B.B., Balasubramanian V.N.* Deep model compression: Distilling knowledge from noisy teachers // arXiv preprint arXiv:1610.09650. 2016.
12. *Furlanello T., Lipton Z., Tschannen M., Itti L., Anandkumar A.* Born again neural networks // International Conference on Machine Learning. PMLR, 2018. С. 1607–1616.
13. *Huang Z., Wang N.* Like what you like: Knowledge distill via neuron selectivity transfer // arXiv preprint arXiv:1707.01219. 2017.
14. *Romero A., Ballas N., Kahou S.E., Chassang A., Gatta C., Bengio Y.* Fitnets: Hints for thin deep nets // arXiv preprint arXiv:1412.6550. 2014.
15. *Chen H., Wang Y., Xu C., Xu C., Tao D.* Learning student networks via feature embedding // IEEE Transactions on Neural Networks and Learning Systems. 2020. Т. 32. No. 1. С. 25–35.
16. *Park W., Kim D., Lu Y., Cho M.* Relational knowledge distillation // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. С. 3967–3976.
17. *Feng Y., Wang H., Hu H.R., Yu L., Wang W., Wang S.* Triplet distillation for deep face recognition // 2020 IEEE International Conference on Image Processing (ICIP). IEEE. 2020. С. 808–812.
18. *Duong C.N., Luu K., Quach K.G., Le N.* Shrinkteanet: Million-scale lightweight face recognition via shrinking teacher-student networks // arXiv preprint arXiv:1905.10620. 2019.
19. *Nekhaev D., Milyaev S., Laptev I.* Margin based knowledge distillation for mobile face recognition // Twelfth International Conference on Machine Vision (ICMV 2019). – International Society for Optics and Photonics, 2020. Т. 11433. С. 114330O.
20. *He K., Zhang X., Ren S., Sun J.* Deep residual learning for image recognition // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. С. 770–778.
21. *Zhang K., Zhang Z., Li Z., Qiao Y.* Joint face detection and alignment using multitask cascaded convolutional networks // IEEE Signal Processing Letters. 2016. Т. 23. № 10. С. 1499–1503.
22. *Guo Y., Zhang L., Hu Y., He X., Gao J.* Ms-celeb-1m: A dataset and benchmark for large-scale face recognition // European conference on computer vision. Springer, Cham, 2016. С. 87–102.

23. *Ng H.W., Winkler S.* A data-driven approach to cleaning large face datasets // 2014 IEEE international conference on image processing (ICIP). IEEE, 2014. С. 343–347.

24. *Robbins H., Monro S.* A stochastic approximation method // The annals of mathematical statistics. 1951. С. 400–407.

25. *Грабовой А.В., Стрижов В.В.* Байесовская дистилляция моделей глубокого обучения // АиТ. 2021. № 11. С. 16–29.

    Grabovoy A.V., Strijov V.V. Bayesian Distillation of Deep Learning Models // Autom. Remote Control. 2021. Т. 82. No. 11. С. 1846–1856.

26. *Грабовой А.В., Стрижов В.В.* Вероятностная интерпретация задачи дистилляции // АиТ. 2022. № 1. С. 150–168.

    Grabovoy A.V., Strijov V.V. Probabilistic Interpretation of the Distillation Problem // Autom. Remote Control. 2022. Т. 83. No. 1. С. 123–137.

27. MarginDistillation: distillation for margin-basedsoftmax: https://github.com/davidsvitov/margindistillation (дата обращения: 08.01.2022).

Статья представлена к публикации членом редколлегии А.А. Лазаревым.

Поступила в редакцию 10.01.2022

После доработки 08.05.2022

Принята к публикации 29.06.2022