

Урок 4

Нейронные сети для анализа изображений

4.1. Классификация изображений

Классификация изображений — это задача, в рамках которой картинку необходимо отнести к одной из нескольких категорий. Примером может служить бинарная классификация изображений на сделанные в помещении и вне помещения, или многоклассовая классификация изображений собак на различные породы.

4.1.1. База ImageNet

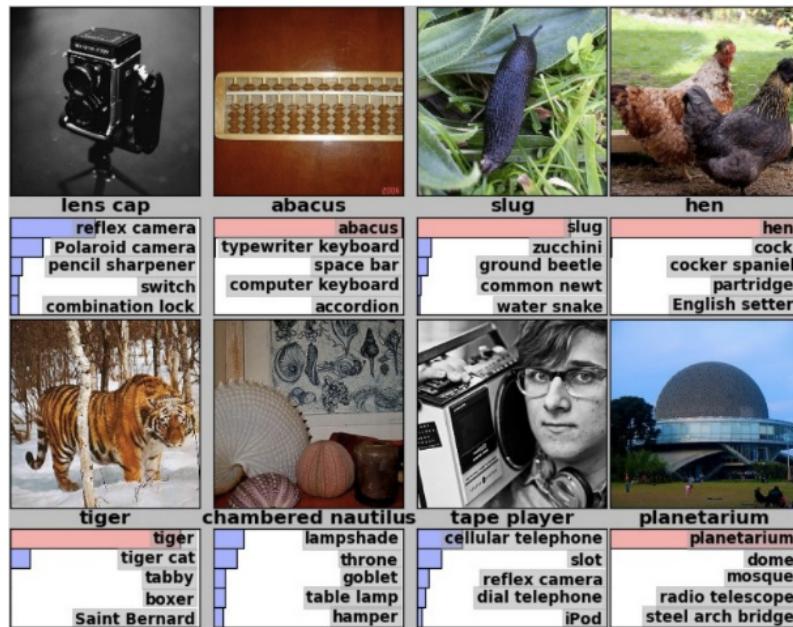


Рис. 4.1: Примеры изображений из базы ImageNet

База ImageNet содержит 10 млн изображений из интернета. Содержащиеся в ней картинки достаточно разнообразные, и все они размечены на принадлежность к тому или иному классу. Примеры изображений из базы показаны на рисунке 4.1.

Стоит отметить, что в базе встречаются ошибки и спорные случаи. Например, изображение на рисунке 4.2 отмечено в базе как "вишня". Если отнести его, например, к классу "дальматинец", то ответ будет считаться неправильным.

На базе ImageNet периодически проводятся соревнования по классификации изображений. Используется 1 млн изображений, каждое из которых требуется отнести к одному из 1000 классов. Можно давать больше



Рис. 4.2: Пример спорной классификации изображения в базе ImageNet, оно отмечено как "вишня"

одного ответа, и если среди первых пяти оказался верный, то классификация считается верной.

До 2012 года лучшие результаты в этом соревновании показывали методы, количество ошибок у которых достигало 25%. Позже лидерство захватили глубокие нейронные сети. Первое их применение показало результат в 16% ошибок, то есть их количество снизилось практически в 2 раза.

4.1.2. Нейронные сети

Можно сказать, что в 2012 году произошла революция в компьютерном зрении. Она началась с того, что с их помощью была одержана победа в соревновании ImageNet. Далее для всей большего количества задач компьютерного зрения было показано, что нейронные сети работают лучше, чем традиционные подходы.

В предыдущих курсах специализации было разобрано, что из себя представляют нейронные сети. Далее упор будет сделан на сети, которые используются для классификации изображений.

Одно из ключевых отличий этих сетей — наличие свёрточных слоёв. Ранее было разобрано, как свёртка используется для обработки изображений. В случай нейронных сетей происходит то же самое, но она применяется не к картинке, а к выходам из предыдущего слоя. Таким образом, каждый слой состоит из банка фильтров.

Помимо свёрточных слоёв в нейронных сетях для классификации изображений применяется операция под названием пулинг. При этом выбирается из нескольких элементов один, а остальные выбрасываются. В частности, тах-пулинг выбирает элемент с максимальным значением.

Другой трюк, который применялся в 2012 году, — это *dropout*. При его использовании обнуляется часть выходов из предыдущего слоя. Можно считать это методом регуляризации, используемом для того, чтобы сеть не переобучалась. Многие считают *dropout* очень неинтуитивным методом регуляризации, тем не менее, на практике показано, что он очень хорошо работает.

Ещё один метод, применяющийся в соревновании ImageNet, — это дополнение обучающих данных. Чем больше выборка, на которой обучается нейронная сеть, тем лучше она впоследствии работает. Исходя из этого, кажется логичным пополнять коллекцию различными способами. В частности, в работе 2012 года применялся метод вырезания кусочка картинки, так, чтобы класс не изменялся. Кроме того, вносились цветовые шумы, производилось зеркальное отражение и производились другие изменения, которые не влияли на содержание

картинки. При обучении на этих данных добавались инвариантности сети к таким изменениям.

4.1.3. AlexNet

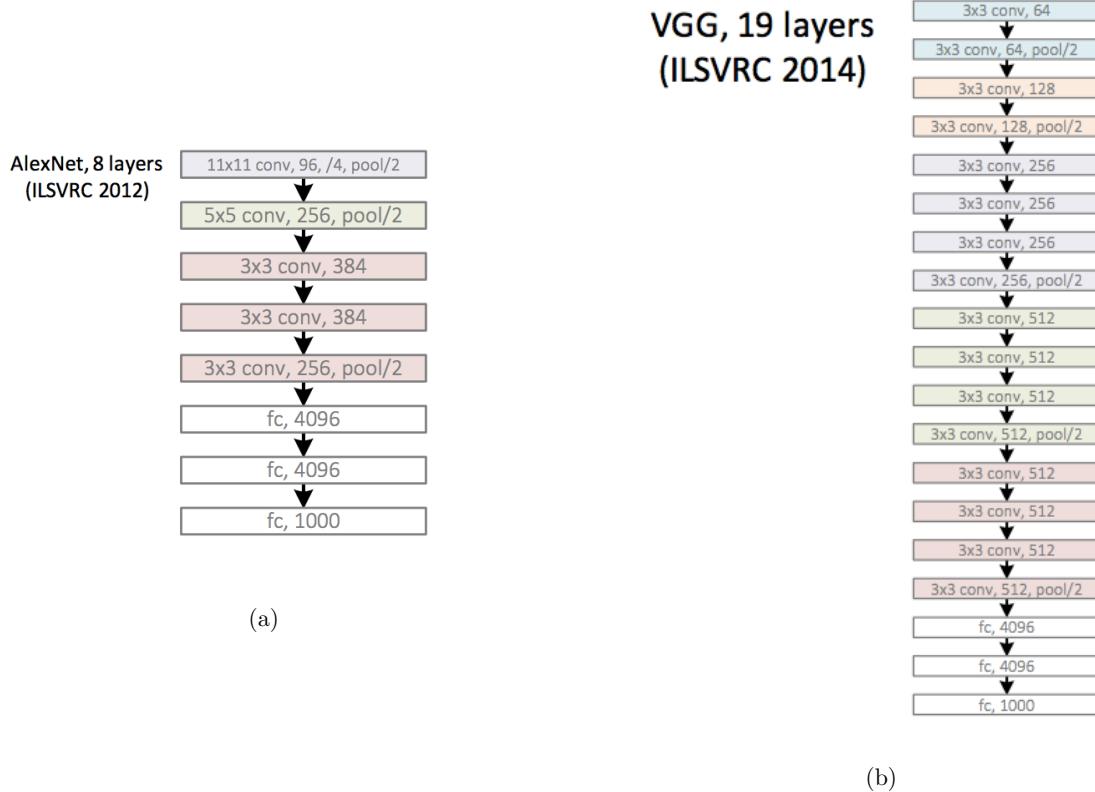


Рис. 4.3: Архитектуры нейронных сетей. (а) — AlexNet, (б) — VGG.

Архитектура знаменитой нейронной сети, победившей в конкурсе в 2012 году, показана на рис. 4.3а. Её разработал и реализовал Алекс Крижевский. Сеть состоит из пяти свёрточных слоёв, двух полносвязных слоёв и выходного слоя с 1000 выходами. Количество выходов совпадает с числом классов в задаче. Таким образом, выход с максимальным значением показывает класс, к которому принадлежит изображение. Данная сеть допустила 16.5% ошибок. Впоследствии возникло много других архитектур, улучшивших этот результат.

4.1.4. Ансамбль сетей

Первое заметное улучшение качества классификации было получено не с помощью изменения архитектуры, а при использовании ансамблей. Идея заключалась в том, чтобы обучить несколько нейронных сетей, каждую картинку классифицировать всеми сетями, а результат усреднить. Такая несложная операция позволила уменьшить долю ошибок с 16.5% до 11.7%.

4.1.5. VGG

В 2014 году группа исследователей из Оксфорда предложила новую архитектуру нейронной сети. Она состояла из 19 слоёв, большая часть из них — свёрточные (все они размера 3×3). Данная сеть позволила улучшить результат классификации и существенно снизить ошибку до 7.3%.

4.1.6. GoogleNet

Ещё одна архитектура — это GoogleNet. Эта нейронная сеть состоит из компонент, показанных на рисунке 4.4. Основная идея состоит в том, что на каждом слое используется не одна свёртка, а несколько, причём

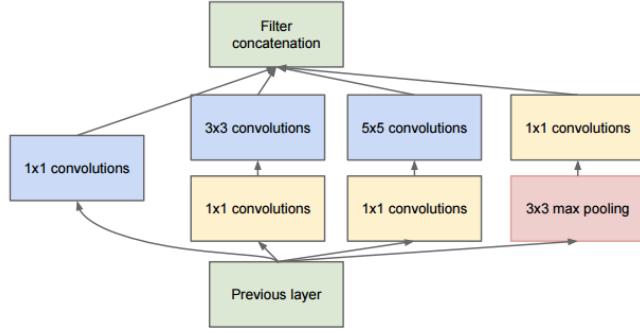


Рис. 4.4: Компоненты свёрток нейронной сети GoogleNet.

разного размера. Это помогает реагировать на сигналы разного масштаба и улучшает качество работы (доля ошибок 6.7%). Полный вид этой сети показан на рисунке 4.5а.

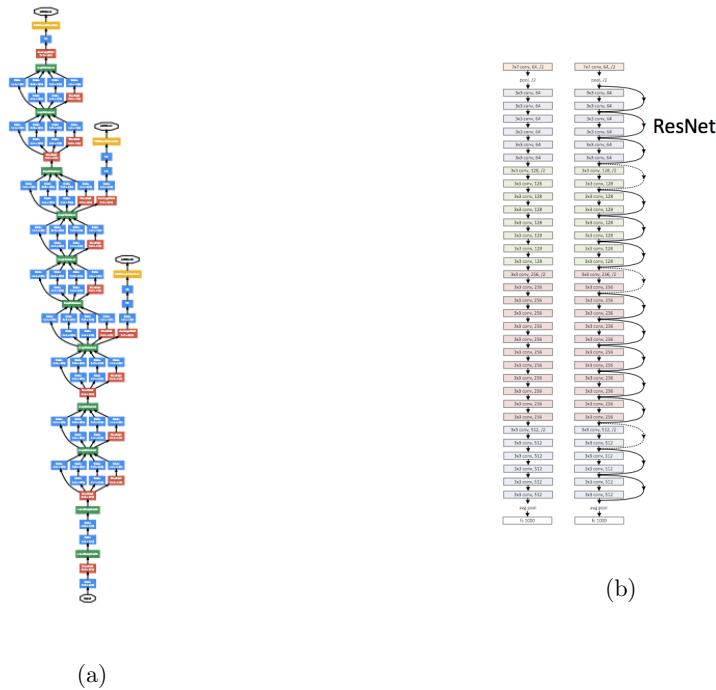


Рис. 4.5: Архитектуры нейронных сетей. (а) — GoogleNet, (б) — ResNet.

4.1.7. ResNet

На рисунке 4.6 показано, как менялись результаты соревнования ImageNet из года в год. Последняя архитектура состоит из 152 слоёв, это Residual Neural Network. С её помощью доля ошибок была уменьшена до 3.57%. Ключевым элементом данной архитектуры является связь, которая пропускает несколько слоёв, передавая результат предыдущего слоя. Такое изменение позволило полностью отказаться от таких методов регуляризации, как DropOut.

Итак, с 2012 года доля ошибок классификации изображений уменьшилась почти в 4 раза, с 16.5% до 3.57%. Это большой скачок, который был осуществлён благодаря развитию нейронных сетей и глубокого обучения. Помимо задачи классификации, нейронные сети применяются для решения практически всех задач компьютерного зрения.

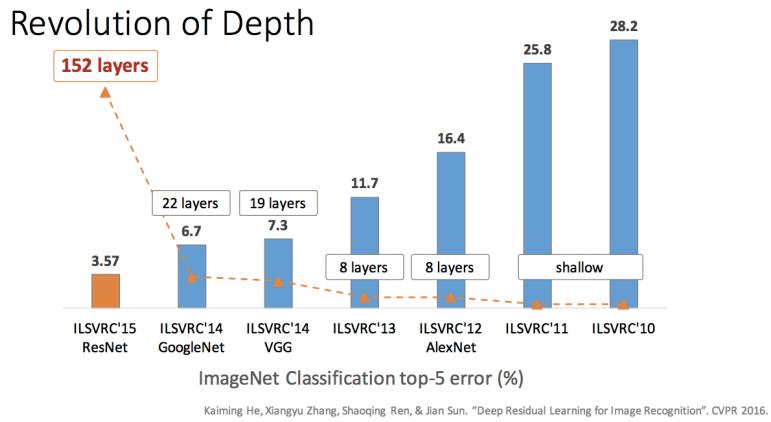


Рис. 4.6: Эволюция ошибки классификации в соревновании ImageNet

4.2. Задача классификации изображений на практике

Далее речь пойдёт о том, как решать задачу классификации изображений на практике.

Прежде всего необходимо собрать базу изображений, для которых известно, к какому классу они принадлежат. Затем нужно определиться с алгоритмом классификации и предобработать картинки. Далее будет разобрано подробнее, как это делать. Последний шаг — это обучение нейронной сети.

4.2.1. Библиотеки для работы с нейросетями

Существует множество библиотек для работы с нейронными сетями. Далее перечислена лишь часть их них:

torch7 написана на C++, всё взаимодействие с клиентским кодом происходит через Lua. Существует достаточно давно, поэтому имеется много реализаций разных задач;

tensorflow — относительно новая, очень удобная библиотека. Минус заключается в том, что из-за новизны в ней часто нет реализаций, которые присутствуют в других библиотеках. С другой стороны, это обычно просто сделать;

theano существует достаточно давно, очень популярна в академических кругах (учёные любят использовать её для иллюстрации статей и быстрых экспериментов). Минус этой библиотеки — её низкоуронвность (требуется много сил и опыта, чтобы понять, как с её помощью работать с сетями), в библиотеке реализованы тензорные вычисления, а не нейронные сети. По этой причине, как правило, работа происходит не с theano, а с обёртками;

keras — обёртка для theano;

lasagne — обёртка для theano;

caffe — одна из первых библиотек для работы со свёрточными нейронными сетями, реализована на C++. Плюс библиотеки: она часто используется для production-задач (потому что написана на C++, а также часть применения нейронных сетей хорошо отлита).

При выборе библиотеки стоит учитывать разнообразные факторы. Например, любовь разработчиков к тому или иному языку программирования (кто-то предпочитает python, кто-то — lua). Также стоит посмотреть, кто уже решал похожую задачу до этого, и в каком фреймворке, и взять эту реализацию за основу. Этот фактор часто является определяющим при выборе библиотеки: если что-то уже реализовано на torch7, зачем использовать что-то другое.

В целом переключаться между этими библиотеками не так сложно, их идеологии достаточно похожи. Поэтому, разобравшись с одной библиотекой, можно воспользоваться и другой. Также существуют конвертеры моделей между фреймворками, что облегчает переход.

4.2.2. Tensorflow

Плюсы библиотеки tensorflow — это богатая документация, большое количество тьюториалов, достаточно простой python api, ядро реализовано на C++.

Часто решение задачи классификации делится на две части. Первая — это подбор модели, обучение классификаторов. Вторая часть — внедрение классификаторов в production. На этом этапе нужна только та часть библиотеки, которая связана с исполнением модели. В некоторых библиотеках (в частности, tensorflow) эту часть легко отделить или можно использовать всю библиотеку, и это никак не повлияет на производительность, а в других — могут возникать проблемы.

4.2.3. Зоопарк моделей

При решении задачи в первую очередь необходимо посмотреть, не решал ли её кто-то до этого, и нет ли готового решения. Для библиотеки caffe существует единый репозиторий моделей (зоопарк моделей, <https://github.com/BVLC/caffe/wiki/Model-Zoo>), можно попробовать искать решение в нём. В частности, там выложены несколько моделей, обученных на базе ImageNet, причём с лучшими результатами на момент обучения. Таким образом, возможно, ничего не нужно придумывать, задача уже кем-то решена. Тогда достаточно взять готовую модель и использовать её на практике.

4.2.4. VGG

Одна из самых часто используемых моделей — это модель VGG, разработанная оксфордской группой исследователей (рис. 4.3b). Её устройство обсуждалось ранее. Реализация этой архитектуры представлена в зоопарке моделей.

Плюсы этой архитектуры заключаются в том, что у неё очень простая структура, она наиболее кроссплатформенна между библиотеками, и, скорее всего, поддерживается любой библиотекой нейронных сетей, которая используется на практике. Кроме того, она показывает достаточно хороший результат на базе ImageNet (не самый лучший, но входит в тройку лидеров).

Существует готовая VGG модель для caffe, кроме того, есть конвертер из этой библиотеки в любую другую.

4.2.5. Дообучение

Итак, стоит задача выполнить классификацию изображений на категории, которых изначально нет в ImageNet. Коллекция уже собрана, теперь есть несколько вариантов развития событий. Можно обучить на этой коллекции нейронную сеть с нуля, как это делалось для задачи ImageNet. Однако сделать это не так просто. Во-первых, коллекция должна быть достаточно большой, а это часто невозможно (разметить миллион картинок — трудоёмкое занятие). Другой вариант — воспользоваться обученной моделью, например, VGG из зоопарка моделей, и дообучить её на собранных данных.

Дообучение можно производить несколькими способами. Во-первых, можно убрать самый последний слой, в котором осуществлялась классификация в ImageNet, и заменить его на новый слой с необходимым количеством классов. Дообучение будет производиться только между этим слоем и последним полносвязным. Плюсы такого подхода заключаются в том, что его достаточно просто осуществить, для этого нужно не так много изображений, не требуется больших мощностей (достаточно среднего ноутбука), сеть впоследствии показывает хорошие результаты. Пример дообучения в библиотеке tensorflow приведён по ссылке (https://www.tensorflow.org/versions/r0.9/how_tos/image_retraining/index.html#distortions).

Другой вариант также предполагает использование готовой, уже обученной на базе ImageNet модели. В ней необходимо заменить последний слой, а затем дообучить все слои, а не только переход от последнего полносвязного слоя к ответу. Для этого требуется больше вычислительных мощностей, но так как сеть уже предобучена, это займёт меньше времени, чем обучение сети с нуля.

Выбор между двумя описанными вариантами зависит от задачи, размера имеющейся базы и желаемого качества классификации. Стоит повторить, что дообучение последнего перехода очень часто даёт хорошие результаты несмотря на то, что это занимает гораздо меньше времени, чем обучение всей сети.

Поиск изображений

Поиск изображений — это ещё одна задача, в которой используются предобученные нейронные сети. Для демонстрации возможного решения этой задачи будет использоваться модель AlexNet (рис. ??), при этом считается, что один из последних слоёв описывает изображение. Это может быть пятый свёрточный слой,

шестой или седьмой полно связанный. При поиске изображения оно будет описываться вектором признаков, взятым из одного из этих слоёв.



Рис. 4.7: Примеры результата поиска похожих изображений. Слева — запрос, остальные картинки — ответы.

Возникает вопрос, как сравнить картинки, чтобы определить, похожи они или нет. Самый простой способ — использовать для этого евклидово расстояние. На рисунке 4.7 показан результат поиска похожих изображений для различных вариантов выбора слоя, описывающего картинку. Видно, что приемлемые результаты получаются, если выбирать в качестве описания седьмой слой: фотографии той же двери расположены на втором и третьем месте. Данный пример является довольно сложным, поэтому результат выходит не очень хорошим. Подробные результаты можно найти по ссылке (<http://sites.skoltech.ru/comppvision/projects/neuralcodes/>). Там представлены подробные соображения о том, какой слой выбирать для описания, как сравнивать вектора признаков, какие комбинации работают лучше в конкретных ситуациях.

4.2.6. Нейронная сеть — хорошее представление

Хочется отметить следующее. Выходы последних слоёв нейронной сети использовались для дообучения классификаторов в разнообразных задачах. Эти же выходы применялись для поиска изображений. Кроме того, нейронная сеть обучалась на коллекции ImageNet, а использовалась для совершенно других наборов изображений. Как показывает практика, эти подходы работают хорошо, даже если модель используется для работы с изображениями абсолютно другой природы (например, с компьютерной графикой, в коллекции ImageNet её нет). Все эти факты говорят о том, что нейронные сети, обученные на большой базе разнообразных изображений, достаточно хорошо описывают картинки. Это фундаментальный результат, показывающий, что данное представление можно использовать в совершенно разных задачах.

4.3. Распознавание лиц

Задачу распознавания лиц можно разделить на две подзадачи. Первая — верификация лиц, требуется по двум фотографиям определить, изображён на них один и тот же человек или нет. Вторая подзадача — распознавание, по фотографии необходимо определить, присутствует ли этот человек в базе (например, базе фотографий злоумышленников), и найти его в ней. Задача распознавания вытекает из задачи верификации, и если научиться хорошо решать первую, то можно решить и вторую. В любом случае, необходимо иметь хорошее представление лиц в сырье виде. Ранее рассказывалось о том, что с помощью нейронных сетей можно получить хорошее представление изображений, остаётся применить это знание.

4.3.1. Базы лиц

Для решения любой задачи машинного обучения (в частности, компьютерного зрения) требуется размеченная база. В мире существует несколько публичных баз лиц. В последнее время чаще всего алгоритмы сравнивают между собой по базе Labeled faces in the wild. Она содержит 13 тыс. фотографий, взятых из интернета. База Megaface содержит 5 млн фотографий, на которых запечатлены 672 тыс. людей. Во второй базе хранится гораздо больше изображений, и они более приближены к реальности. Однако она появилась недавно, не все алгоритмы были на ней протестированы, поэтому для сравнения качества работы всё ещё используется первая база.

4.3.2. Пары лиц



(a)

(b)

Рис. 4.8

Чтобы оценить сложность задачи верификации, полезно рассмотреть пример. На рисунке 4.8 представлены две пары изображений, на одной из них разные люди, а на другой — один и тот же человек. На самом деле, не все люди могут сразу определить, на какой паре изображён один и тот же человек. Плохое качество изображений связано не с желанием усложнить задачу, оно такое изначально, потому что лица людей, вырезанные с фотографий из интернета, получаются маленькими. Фотографии такого же качества даются на вход нейронной сети.

4.3.3. Фреймворк распознавания

Почти все алгоритмы классификации и распознавания лиц можно разделить на несколько фаз. Сначала необходимо найти на фотографии лицо и вырезать. Затем его нужно развернуть таким образом, чтобы оно имело примерно одно положение на всех фотографиях. Это упрощает работу алгоритмов классификации. Из полученной картинки генерируется цифровое представление лица. Последний шаг — сравнение полученных лиц. Далее каждый из этих этапов будет рассмотрен подробнее.

Детектор

Ранее упоминалось, что существует много методов детекции лица. Один из первых успешных использует каскады Хаара, сейчас это не самый лучший и успешный метод. В данный момент для решения этой задачи используются нейронные сети. Кроме того, достаточно хорошим компромиссом между качеством, скоростью и удобством использования является следующий метод. Из изображения извлекаются дескрипторы hog, а затем по ним алгоритм машинного обучения svm определяет, лицо это или нет.

Существуют методы выделения особенностей лица, таких как кончик носа, уголки рта, края глаз и т.д. Зная эти особые точки, можно тем или иным способом развернуть лицо.

Выравнивание

Существует несколько способов выравнивания лица. Можно развернуть изображение в плоскости, выполнив двумерное преобразование. Другой способ — попытаться построить трёхмерную форму лица, и сделать так, будто человек смотрит прямо в камеру.

До появления нейронных сетей качество выравнивания было одним из ключевых параметров алгоритма распознавания лиц. В соревнованиях побеждали методы, создающие качественное 3D-представление, разворачивающие лицо, дорисовывающие вторую половину лица, если на фото изображён профиль. Однако нейронные сети могут хорошо работать даже на не до конца выравненных лицах, поэтому требования к этому этапу работы алгоритма снизились. Более того, простые методы (например, поворот в плоскости) даже выигрывают, потому что сложные методы выравнивания вносят шум в изображение.

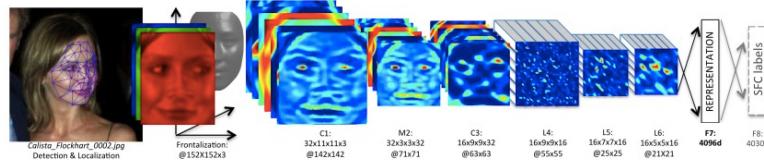


Рис. 4.9: Архитектура нейронной сети, использующаяся для классификации лиц (классы — люди).

Обучение представления

После того как лицо найдено и выравнено, можно перейти к работе непосредственно с картинками. Для этого применяется уже не раз упоминавшаяся схема классификации изображений с использованием глубоких нейронных сетей. На рисунке 4.9 показана похожая на AlexNet архитектура, которая применяется для классификации лиц в статье. Эта нейронная сеть состоит из нескольких свёрточных и полно связанных слоёв, она обучалась на базе, состоящей из нескольких миллионов фотографий, в роли классов выступают различные люди. После обучения сети выходной слой использовался в качестве описания картинки, с помощью которого решалась задача верификации на базе Labeled faces in the wild.

Возникает вопрос, как сравнивать между собой представления изображений. До появления нейронных сетей это было важным компонентом всей цепочки метода, и были придуманы различные способы производить сравнение. В случае нейронных сетей хорошо работает обычное евклидово расстояние: если оно больше заданного порога, то люди разные. Сложные методы либо вообще не дают никакого улучшения в этом случае, либо это улучшение совсем небольшое.

В статье, использующей описанную выше архитектуру, была достигнута точность 97.35%. Это первая работа на базе нейронных сетей, которая значительно улучшила результат (ошибка снизилась в несколько раз, точность используемых ранее методов составляла около 92%).

Ансамбль сетей — deep id

Далее авторы из Гонконга предложили использовать похожий пайплайн и похожую сеть, они добились увеличения точности до 99.47%. Различие заключалось в том, что для этого использовалось 200 нейронных сетей, результаты которых специальным образом объединялись и усреднялись. Стоит отметить, что даже если нейронная сеть небольшая, применять её 200 раз — сложная процедура.

Оксфорд

Ещё один хороший результат был получен группой из Оксфорда, при этом использовалась одна нейронная сеть. Они изменили протокол обучения, база лиц была относительно небольшой (2 млн картинок). При этом было получено качество 98.95%. Впоследствии результат был улучшен до 99.13%, для этого было изменено обучение сравнения дескрипторов.

End-to-end

Как уже было сказано ранее, фреймворк распознавания состоит из детектирования лица, его выравнивания и обучения нейронной сети. Сотрудники компании Google решили, что все эти шаги могут выполняться внутри сети. Они использовали базу из 200 млн изображений и без предобработки обучили на ней сеть. Полученное качество составляло 98.87%. Кажется, что большой размер базы должен способствовать тому, чтобы нейронная сеть сама научилась выравнивать изображения. Однако даже в этом случае предварительное выравнивание сильно помогает, с его использованием качество достигло 99.63%.

4.3.4. Результаты

В таблице 4.1 суммированы результаты верификации лиц, полученные с использованием различных методов. Важно учитывать не только качество работы метода, но и размер базы, на которой производилось обучение: чем больше база, тем сложнее её собрать. Например, поражает воображение точность, полученная при обучении на базе из 200 млн изображений, но повторить его достаточно сложно. С другой стороны, группа из Оксфорда показала очень достойный результат, используя относительно небольшую базу.

№	Метод	Число изображений, млн	Количество сетей	Точность, %
1	Fisher Vector Faces	-	-	93.10
2	DeepFace	4	3	97.35
3	Fusion	500	5	98.37
4	DeepID-2,3		200	99.47
5	FaceNet	200	1	98.87
6	FaceNet + Alignment	200	1	99.63
7	Ours	2.6	1	98.95

Таблица 4.1: Результаты верификации лиц различными методами

Таким образом можно считать, что на базе Labeled faces in the wild, состоящей из 13 тыс. изображений, задача распознавания решена достаточно хорошо, с точностью почти 100%.

4.3.5. Поиск похожих лиц

Ещё один пример использования представления, полученного при использовании классификаторов, — это поиск похожих лиц. Так же как в задаче поиска похожих изображений, для этого можно использовать несколько последних слоёв нейронной сети. В случае, если в базе уже есть данный человек, будет найден он, иначе — просто похожие люди. В данный момент существует несколько сервисов, работающих по этому принципу.

4.3.6. t-sne

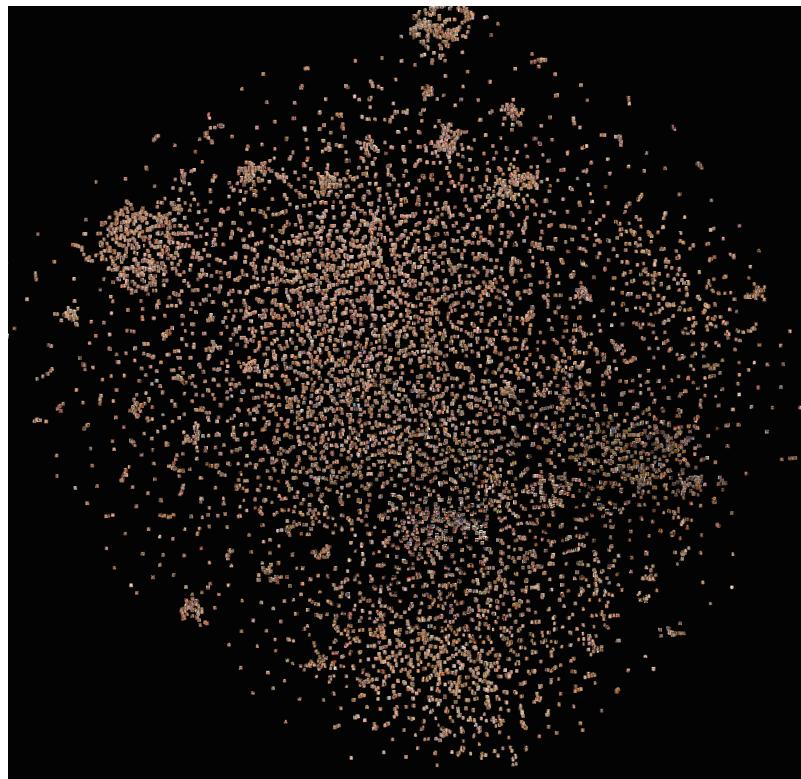


Рис. 4.10: Визуализация на плоскости признаковых описаний лиц, полученных с помощью нейронных сетей.

Чтобы лучше понять, как устроены представления лиц с помощью нейронных сетей, можно визуализировать облако векторов на плоскости (4.10). Для этого можно использовать метод t-SNE, который располагает вектора на плоскости таким образом, чтобы они были тем ближе, чем ближе в исходном пространстве. На рисунке есть несколько сгущений, можно их приблизить и посмотреть, что они из себя представляют (рис.

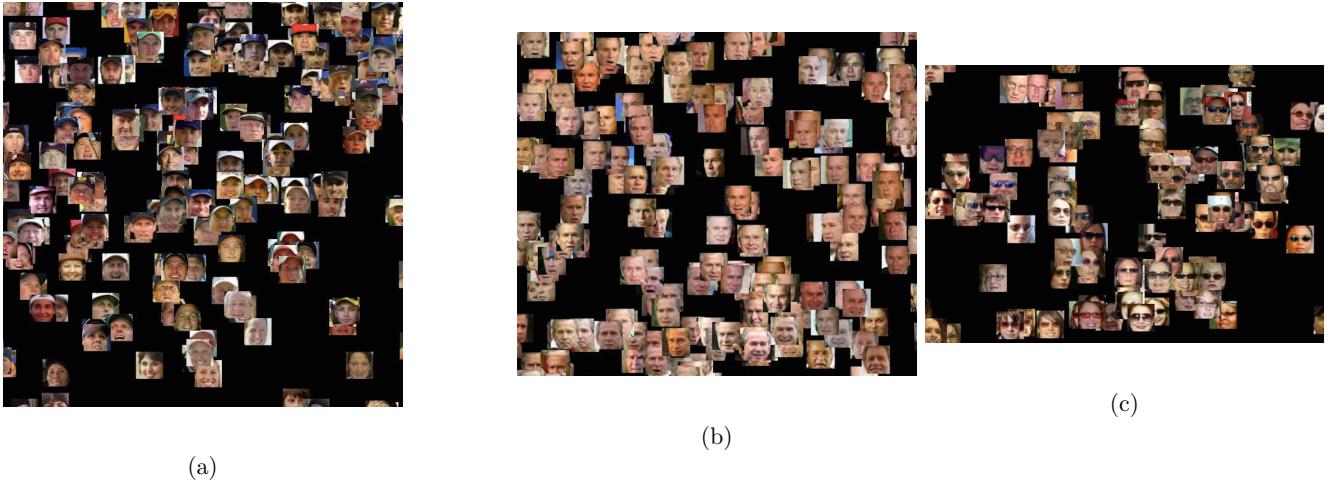


Рис. 4.11: Сгущения из рисунка 4.10. (а) — люди в кепках, (б) — Джордж Буш, (с) — люди в очках

4.11). Интересно, что данные сгущения никак не помогают классифицировать людей. Например, тёмные очки (рис. 4.11с) могут присутствовать у разных людей. Тем не менее, нейронная сеть обратила на это внимание, и считает людей в солнцезащитных очках похожими. Это означает, что полученные признаки неидеальны, и, возможно, произошло переобучение.

4.3.7. Резюме

За последние несколько лет в задаче распознавания лиц произошёл большой прогресс, многократно увеличилось качество работы алгоритмов на одной из баз. На базе Labeled faces in the wild получены результаты с точностью почти 100%, это означает, данная коллекция исчерпала себя. Цель на ближайшие годы для исследователей в этой области — добиться хорошего качества на базе Megaface, содержащей большее количество разнообразных изображений. Если эта цель будет достигнута, это означает, что алгоритмы компьютерного зрения будут хорошо работать на огромных коллекциях (социальные сети, изображения с уличных камер и камер в метро).