УДК 004.89

ОСНОВОПОЛАГАЮЩАЯ КОНЦЕПЦИЯ ТЕХНОЛОГИЧЕСКОГО ПРОРЫВА ГЛУБОКОЙ НЕЙРОННОЙ СЕТИ ALEXNET В ЗАДАЧЕ РАСПОЗНАВАНИЯ ИЗОБРАЖЕНИЙ

Майтак Р. В., бакалавр

Научный руководитель: Дягилева А. В., к.т.н., доцент

Кузбасский государственный технический университет
имени Т.Ф. Горбачева

Перед тем, как приступить к аналитике предметной области распознавания изображений, отметим, что в теории прикладного искусственного интеллекта нет понятия «нейронные сети для анализа изображений» [1].

Поскольку каждое изображение в памяти компьютера представляется в бинарном виде, то, соответственно, для кодирования в дихотомном виде такой информации необходимо будет выделить гораздо больше места, чем, например, для обычного текстового файла. По этой причине, для анализа подобной информации в теории искусственного интеллекта выделено направление глубоких нейронных сетей [2].

Глубокие нейронные сети в классическом понимании подразделяются на два типа [2, 3]:

1. Сверточные нейронные сети;
2. Рекуррентные нейронные сети.

Первый тип глубоких нейронных сетей используется в обработке изображений, в то время как второй чаще применяется для исследования сигналов и текстов (обработка естественного языка).

В задаче классификации изображений изначально был принят подход, основанный на предварительной обработке изображений (рисунок 1).

Рисунок 1 – Классический подход аналитики изображений

Из рисунка 6 следует, что, например, при обработке фотографий людей, первоначально необходимо выделять овал лица, глаз, измерять расстояние между глазами и так далее. Общая совокупность признаков может достигать нескольких десятков или даже сотен признаков для каждого отдельного экземпляра (лица).

Разумеется, что при классическом подходе исследуемая задача становится крайне специфичной и специализированной. То есть, даже в том случае, если модель будет идеально обучена для распознавания лиц, то, например, применить её же в сфере распознавания рукописных цифр будет уже нельзя.

Таким образом, классический подход предполагает ручное создание признаков, основанных на геометрических формах каких-либо изображений. Для каждой специфической задачи будет своя система признаковых описаний, что не позволяет осуществлять переход разработанной модели между различными предметными областями даже в тех случаях, если эти отличия минимальны.

В связи с высокой сложностью работы над каждой отдельной задачей в прикладном искусственном интеллекте возник спрос на концепцию, которая бы подразумевала выполнение процедуры экстракции признаков одновременно с главной задачей нейронной сети – распознаванием и прогнозированием целевого класса [4].

Такой концепцией стал современный подход – «end-to-end deep learning», который предполагает (рисунок 2) автоматическое построение прецедентов по каждому элементу обучения, вне зависимости от наличия у элемента классических векторных признаковых описаний.

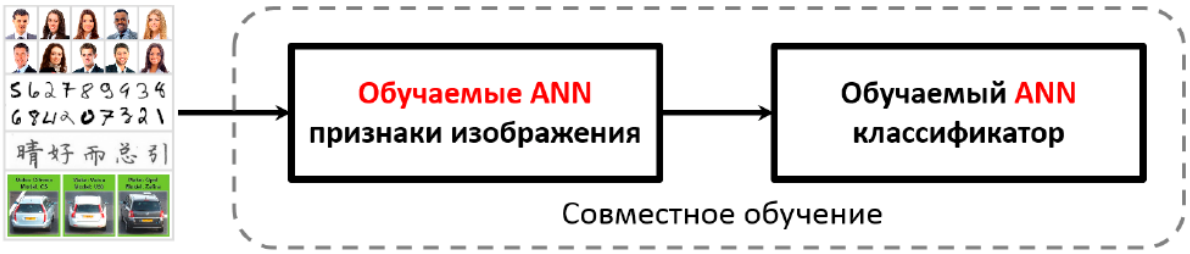

Рисунок 2 – Современный подход аналитики изображений

Изображение является ярким примером такого элемента – это очень большая матрица пикселей. Несмотря на то, что каждый пиксель может быть описан в палитре трех цветов RGB [4], при развертывании всего изображения в векторе в этом признаковом описании потеряется топология картинки (наличия соседних элементов, их цветов, схем и так далее).

Рассмотрим, каким образом для изображений осуществляется предобработка, которая оказывается универсальной для любых изображений: лиц, снимков МРТ, иероглифов, автомобилей и любых других графических объектов.

Для операции предобработки используется свертка. Именно по названию этой основополагающей операции нейронные сети и получили свое именование.

В формате формализованного математического языка $x[i,j]$ – это исходные признаки (пиксели) изображения. w_{ab} – это ядро свертки. Тогда, чтобы получить **(1)** сглаженное изображение (свертку) $(x * w)[i,j]$ в центре некоторой окрестности, необходимо вычислить линейную комбинацию, обойдя всю окрестность с размерами $A \times B$.

$$(x * w)[i,j] = \sum_{a=-A}^{A} \sum_{b=-B}^{B} w_{ab} \times [i+a, \quad j+b] \quad (\mathbf{1})$$

Графически формулу **(1)** можно представить в виде визуального представления математической операции (рисунок 3).

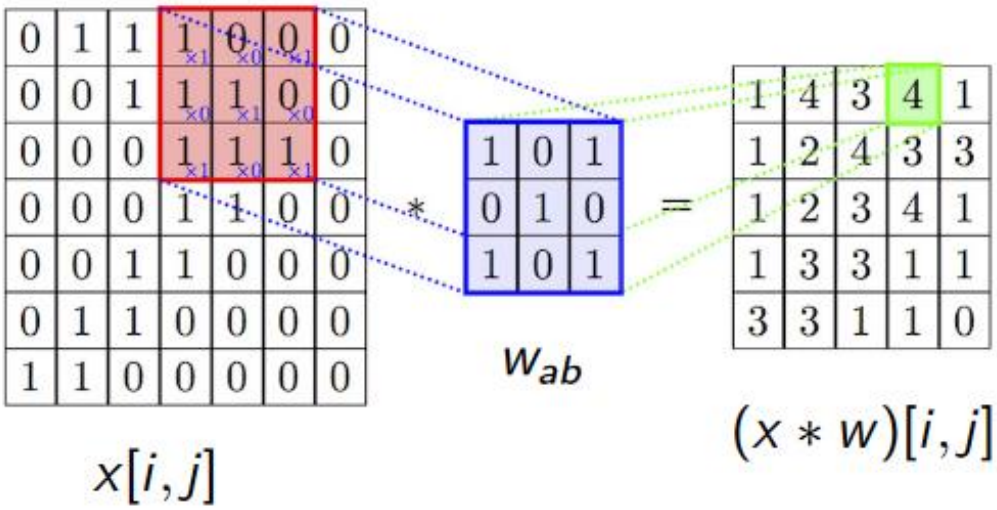

Рисунок 3 – Операция свертки в сверточных нейронных сетях

Формула **(1)** является частным случаем сглаживания. Непосредственно методика сглаживания была известна уже очень давно и перешла в задачи глубокого обучения из других предметных областей:

- Решение задачи шумоподавления;
- Формула Надарая-Ватсона при решении задачи непараметрического восстановления регрессии [3];
- Ядерное сглаживание.

В ядерном сглаживании используется та же самая структура. Единственное отличие формулы **(1)** от ядерного сглаживания состоит в том, что в данном случае **(1)** используется многомерное сглаживание, а не одномерная структура временных рядов.

Кроме того, формула **(1)** представляет собой упрощенную структуру линейного нейрона, в котором отсутствует функция активации [4].

Рассмотрим другой вариант нейрона, который может быть использован в свертке – объединяющий нейрон.

Объединяющий нейрон не имеет обучающихся весовых коэффициентов, его задача сводится к агрегации значений, которые он «видит» в данной окрестности (рисунок 4).

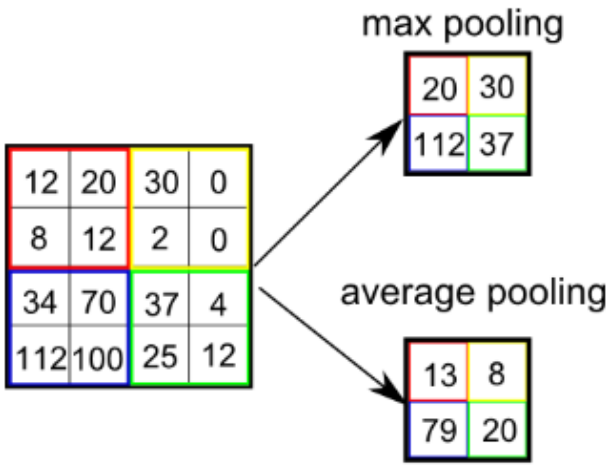

Рисунок 4 – Объединяющий нейрон

Например, на рисунке 4, объединяющий нейрон применяет к каждой ячейке агрегирующую функцию F (она может быть максимизирующей – max pooling, средней – average pooling и любой другой, выполняющей несложные математические операции).

Размер ячейки для свертки задан как 2*2 пикселя, но сам по себе он зависит только от исследователя данных. Свертка позволяет снижать размерность ячейки (агрегирующая функция F «сворачивает» ячейку 2*2 пикселя до размера 1*1 пиксель). В первом случае в качестве функции выбрана

максимизирующая, поэтому в качестве единственного значения ячейки 1*1 остается максимальное значение первоначальной ячейки 2*2. Во втором случае функция F является функцией среднего, поэтому единственным значением пикселя остается среднее значение четырех значений.

В анализе изображений алгоритмами глубокого обучения часто используется чередование метода свертки и max pooling (выбора максимального значения). Смысл этой методики состоит в том, что сверточный слой выполняет усреднение, при этом практически не снижая размеры самого изображения. А слой max pooling позволяет уменьшить размеры самого изображения в кратное число раз. То есть, если изначально размер окрестности был h, то на выходе размер изображения уменьшится в h раз **(2)**.

$$y[i,y] = F\left(x\left[h_i, h_j\right], \dots, \times \left[h_i + h - 1, h_j + h - 1\right]\right) \quad (\mathbf{2})$$

Если сверточный слой научился распознавать определенный элемент изображения, то на выходе он будет выдавать большую величину значения, сигнализируя о определении на изображении объекта из целевого класса. Ценность слоя max pooling состоит в том, что он позволяет определять такие элементы на изображении вне зависимости от места их дислокации на нем.

Для лучшего понимания, рассмотрим стандартную многослойную схему сверточной нейронной сети (от английского «convolutional neural network»), представленную на рисунке 5.

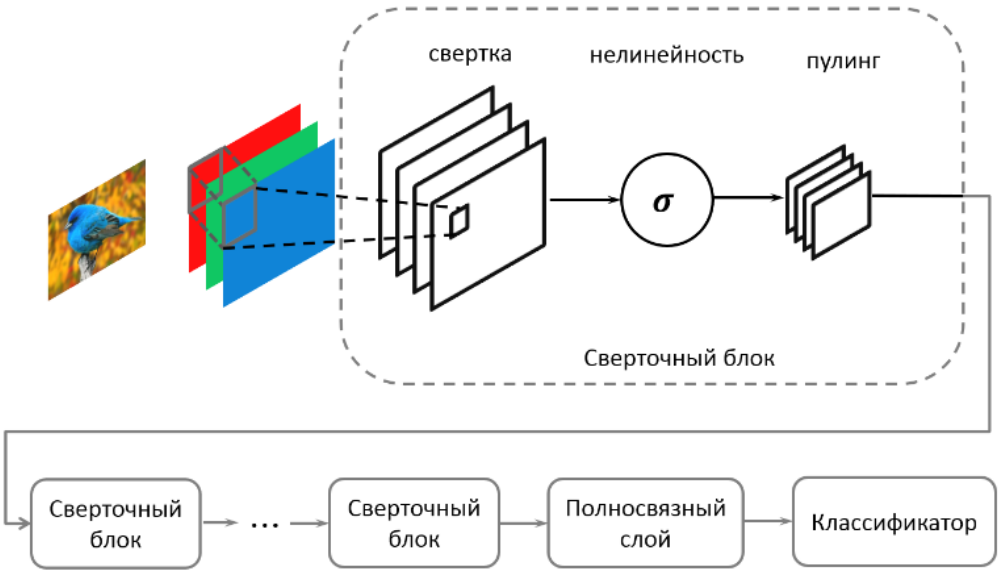

Рисунок 5 – Стандартная схема конфигурации сверточной нейронной сети

В каждом блоке сверточной нейронной сети (рисунок 5) есть два основных слоя – свертка и «пулинг» (от английского «pooling» – «объединяющий нейрон»). Сверточные блоки повторяются друг за другом и эффектом каждого блока становится уменьшение размеров изображения на выходе.

Каждый сверточный нейрон применяется к определенному пикселю входного изображения, поэтому для аналитики изображения требуется большое количество нейронов.

Рассмотрим пример (рисунок 6) сверточной нейронной сети, которая называется AlexNet [3]. В 2012 году именно эта конфигурация нейронной сети и осуществила переворот в распознавании изображений, так как была видоизменена структура формирования сверточных слоев.

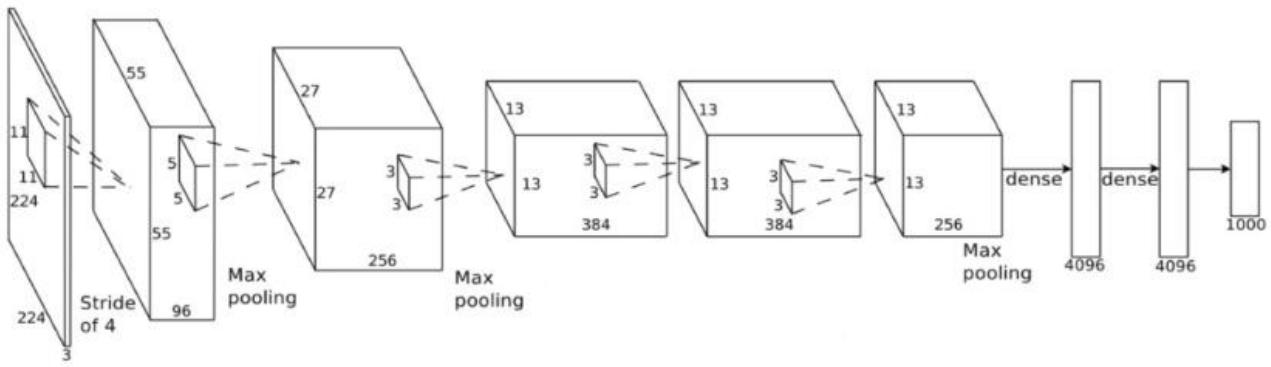

Рисунок 6 – Сверточная нейронная сеть AlexNet

Чтобы лучше понять краеугольную методику функционирования сети AlexNet, рассмотрим методологию аналитики изображения в ней на примере рисунка 11. Исследовать рисунок будем по ходу анализа изображения сетью (то есть, перемещаясь слева направо).

Самым первым элементом сети является непосредственно изображение (размерами 224*224 пикселя). Значение «3» характеризует факт определения цветов в изображении палитрой RGB (размерность вектора каждого пикселя).

Размерность окрестности свертки составляет 11*11.

На втором блоке появляется число 96. Оно определяет количество различных сверток, которые вычисляются в каждом пикселе изображения.

Далее, слой за слоем, благодаря пулингу, уменьшается размерность изображения. В конечном счете, изображение становится стянутым до размеров 1*1 пиксель.

Размерность изображения уменьшается в каждой свертке, но, чтобы не терять данные, соизмеримо увеличивается размерность вектора каждого цвета.

Таким образом, постоянно уменьшая размерность самого изображения, была получена его векторизованная форма. Иными словами, *в автоматическом режиме* были сгенерированы признаки изображения (так как не теряются взаимосвязи между соседними пикселями, их расположением, цветом и так далее).

После векторизации изображения, вектор можно использовать в любых доступных исследователю моделях. Например, добавив всего 2 – 3 полносвязных нейронных слоя (которые уже решают задачу классификации), можно получить нейронную сеть для распознавания объектов на изображении.

Отметим, что такая сеть состоит из двух больших, условно независимых друг от друга, но сильно влияющих на результат по отдельности, частей:

1. Сверточные слои для выполнения векторизации изображения;
2. Малослойная полносвязная нейронная сеть для распознавания изображений.

Результат работы каждого слоя нейронной сети (рисунок 6) можно условно интерпретировать (рисунок 7).

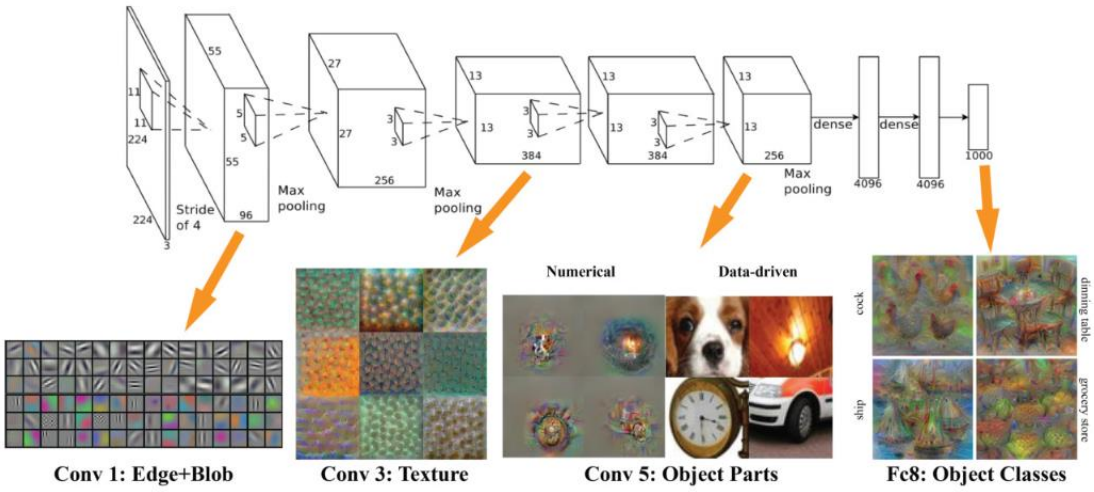

Рисунок 7 – Интерпретация результатов работы нейронной сети

На рисунке 7 показано, что «делает» каждый слой сети. Перебирая изображения для каждого слоя, можно найти такие, которые приводят к высокому отклику нейронов в каждом конкретном слое нейронной сети. Обнаружив, на

какие типы изображения появляется сильный отклик слоя, можно понять, что данный слой распознает.

Оказывается, что первые слои распознают переходы градиентов на изображении, грани и переходы. Как было сказано выше, некоторые свертки могут использовать операции агрегирования среднего, а поскольку операции выполняются над RGB-векторами изображения, то это объясняет, почему некоторые грани в первом слое являются цветными, а некоторые черно-белыми.

Далее, следуя слой за слоем можно наблюдать как слой max pooling позволяет уменьшать размер картинки, в следствии чего каждый слой «видит» более крупные элементы изображений, чем видели предыдущие слои.

На следующем слое (втором) можно заметить распознанные текстуры. Более дальние слои распознают более сложные, но пока не интерпретируемые объекты изображений. Чем дальше переходим к следующим слоям, тем более интерпретируемыми становятся сами объекты (элементы) изображений.

Из этого наблюдения следует логичный вывод, который подтверждается результатами наблюдений рисунка 7: чем дальше по сети (слева направо) расположен слой, тем более крупные и сложные элементы изображений он способен распознавать [3, 4].

Список используемой литературы:

1. Томас Кормен, Чарльз Лейзерсон. Алгоритмы: построение и анализ, 3-е издание – М.: ООО И.Д. Вильямс. 2013. – 1328 с.
2. Yinyan Zhang. Deep Reinforcement Learning with Guaranteed Performance. – Springer Press. 2020. – 265 с.
3. Дж. Клейнберг, Е. Тардос. Алгоритмы: Разработка и применение – СПб.: Питер. 2016 – 800 с.
4. Майкл Солтис. Введение в анализ алгоритмов – СПб.: ДМК. 2019. – 269 с.