

Машинное обучение (Machine Learning)

Глубокое обучение: Сверточные сети (Deep Learning: Convolutional nets)

Уткин Л.В.

Санкт-Петербургский политехнический университет Петра Великого



Содержание

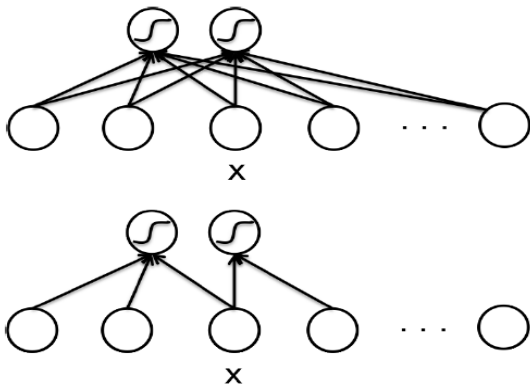
- 1 Причины создания сверточных сетей и предварительные замечания
- 2 Сверточный слой
- 3 Max-pooling
- 4 Некоторые особенности и детали сверточных сетей

Причины создания сверточных сетей и предварительные замечания

Причины создания сверточных сетей

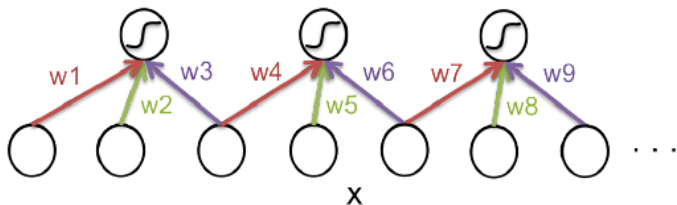
- Вход - большой размерности: каждый нейрон имеет огромное число соединений. Малая картинка 100x100 пикселей (размерность входа - 10000), каждый нейрон имеет 10000 параметров. Если скрытый слой - 2000 нейронов, то всего 2×10^7 соединений.
- А что, если часть соединений убрать?

Две сети



Как уменьшить число соединений?

- Сделать часть весов одинаковыми ("weight sharing" или свертка)
- $w_1 = w_4 = w_7$, $w_2 = w_5 = w_8$, $w_3 = w_6 = w_9$ ("фильтр")
- Вместо хранения всех весов, храним w_1 , w_2 , w_3



Сверточные сети

К свертке добавляется другой слой, известный как “**max-pooling layer**”, который вычисляет максимальное значение из выбранного подмножества выходных нейронов сверточного слоя и использует его как входное значение следующего слоя

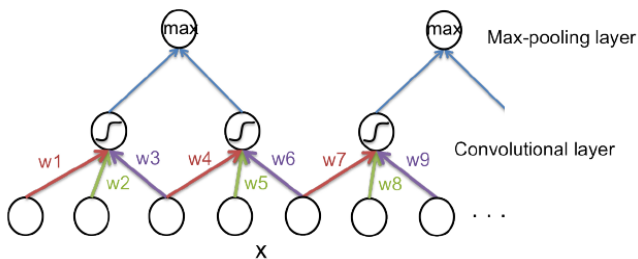
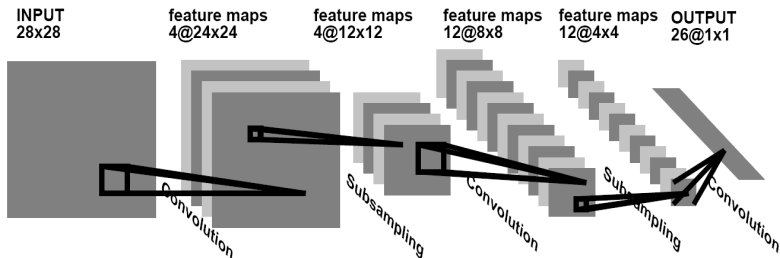


Иллюстрация сверточной сети

LeCun Y., Bengio Y. Convolutional Networks for Images, Speech, and Time-Series // The Handbook of Brain Theory and Neural Networks, MIT Press, 1995

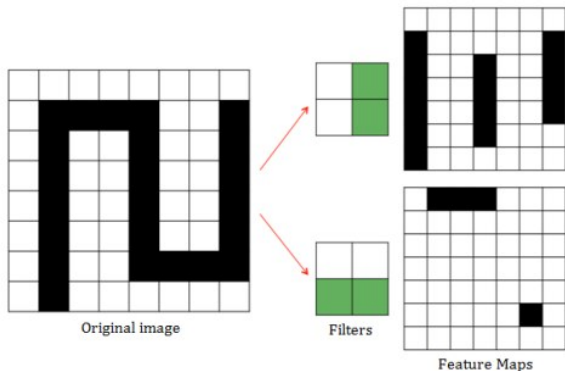


Слои сверточной сети

- **Сверточный слой:** реализует обычную операцию свертки - двигаясь по изображению скользящим окном, перемножаем значения в окне с заданными весами (ядром), а затем все складываем. Наборов весов может быть несколько.
- **Pooling Layers:** Для уменьшения числа данных и выбора наиболее интересных признаков. Входной двумерный массив делится на сектора, в зависимости от параметров, и в каждом из них происходит максимизация (MAX) или усреднение (AVE) (два самых распространенных вида pooling'a).

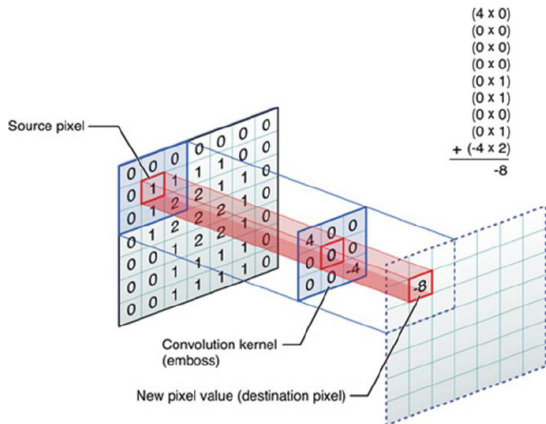
Сверточный слой

Сверточный слой (фильтр)



Цель - определение горизонтальных и вертикальных границ. Используем фильтры (зеленый цвет) – это небольшая матрица, представляющая признак, который хотим найти на исходном изображении.

Операция свертки



Padding

- Пиксели, которые находятся на границе изображения участвуют в меньшем количестве сверток, чем внутренние.
- Поэтому в сверточных слоях используется дополнение изображения (англ. padding).
- Выходы с предыдущего слоя дополняются пикселями так, чтобы после свертки сохранился размер изображения. Такие свертки называют одинаковыми (same convolution), а свертки без дополнения изображения называются правильными (valid convolution).

Padding

0	0	0	0	0	0	0
0	3	2	1	0	0	0
0	0	0	1	3	1	0
0	3	1	2	2	3	0
0	2	0	0	2	2	0
0	2	0	0	0	1	0
0	0	0	0	0	0	0

6.0	0.0	0.0
8.0	0.0	0.0
6.0	4.0	4.0

0	0	0	0	0	0	0
0	3	2	1	0	0	0
0	0	0	1	3	1	0
0	3	1	2	2	3	0
0	2	0	0	2	2	0
0	2	0	0	0	1	0
0	0	0	0	0	0	0

6.0	0.0	0.0
8.0	0.0	0.0
6.0	4.0	4.0

0	0	0	0	0	0	0
0	3	2	1	0	0	0
0	0	0	1	3	1	0
0	3	1	2	2	3	0
0	2	0	0	2	2	0
0	2	0	0	0	1	0
0	0	0	0	0	0	0

6.0	0.0	0.0
8.0	0.0	0.0
6.0	4.0	4.0

0	0	0	0	0	0	0
0	3	3	2	1	0	0
0	0	0	1	3	1	0
0	3	1	2	2	3	0
0	2	0	0	2	2	0
0	2	0	0	0	1	0
0	0	0	0	0	0	0

6.0	0.0	0.0
8.0	0.0	0.0
6.0	4.0	4.0

0	0	0	0	0	0	0
0	3	3	2	1	0	0
0	0	0	1	3	1	0
0	3	1	2	2	3	0
0	2	0	0	2	2	0
0	2	0	0	0	1	0
0	0	0	0	0	0	0

6.0	0.0	0.0
8.0	0.0	0.0
6.0	4.0	4.0

0	0	0	0	0	0	0
0	3	3	2	1	0	0
0	0	0	1	3	1	0
0	3	1	2	2	3	0
0	2	0	0	2	2	0
0	2	0	0	0	1	0
0	0	0	0	0	0	0

6.0	0.0	0.0
8.0	0.0	0.0
6.0	4.0	4.0

0	0	0	0	0	0	0
0	3	3	2	1	0	0
0	0	0	1	3	1	0
0	3	1	2	2	3	0
0	2	0	0	2	2	0
0	2	0	0	0	1	0
0	0	0	0	0	0	0

6.0	0.0	0.0
8.0	0.0	0.0
6.0	4.0	4.0

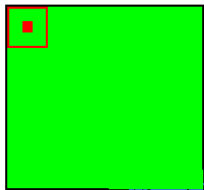
0	0	0	0	0	0	0
0	3	3	2	1	0	0
0	0	0	1	3	1	0
0	3	1	2	2	3	0
0	2	0	0	2	2	0
0	2	0	0	0	1	0
0	0	0	0	0	0	0

6.0	0.0	0.0
8.0	0.0	0.0
6.0	4.0	4.0

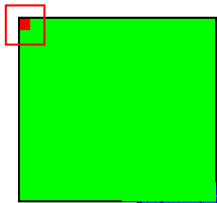
0	0	0	0	0	0	0
0	3	3	2	1	0	0
0	0	0	1	3	1	0
0	3	1	2	2	3	0
0	2	0	0	2	2	0
0	2	0	0	0	1	0
0	0	0	0	0	0	0

6.0	0.0	0.0
8.0	0.0	0.0
6.0	4.0	4.0

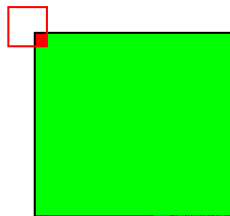
Padding



valid



same



full

Результат операция свертки



-1	0	+1
-2	0	+2
-1	0	+1

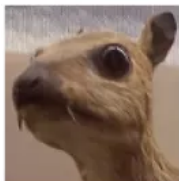
Gx

+1	+2	+1
0	0	0
-1	-2	-1

Gy

Сверточный слой (фильтр)

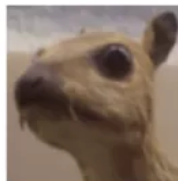
Input image



Kernel

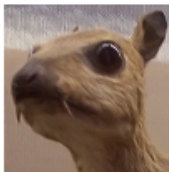
$$\begin{pmatrix} \frac{1}{16} & \frac{1}{8} & \frac{1}{16} \\ \frac{1}{8} & \frac{1}{4} & \frac{1}{8} \\ \frac{1}{16} & \frac{1}{8} & \frac{1}{16} \end{pmatrix}$$

Feature map



Сверточный слой (фильтр)

Input image




Convolution
Kernel

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$




Feature map






Сверточный слой (фильтр)

Operation	Filter	Convolved Image
Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	

Сверточный слой (фильтр)

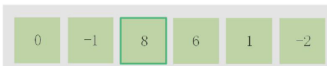
Operation	Filter	Convolved Image
Edge detection	$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$	
	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$	
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	

Сверточный слой (фильтр)

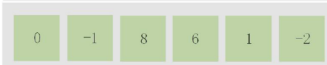
Operation	Filter	Convolved Image
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Box blur (normalized)	$\frac{1}{9} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	
Gaussian blur (approximation)	$\frac{1}{16} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$	

Max-pooling

Max-pooling



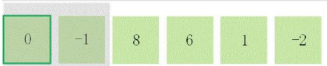
Mean-pooling



k-max-pooling



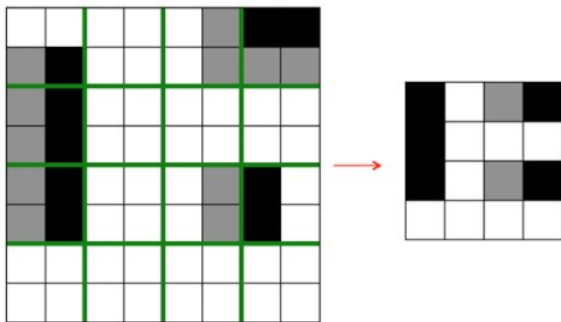
Dynamic pooling



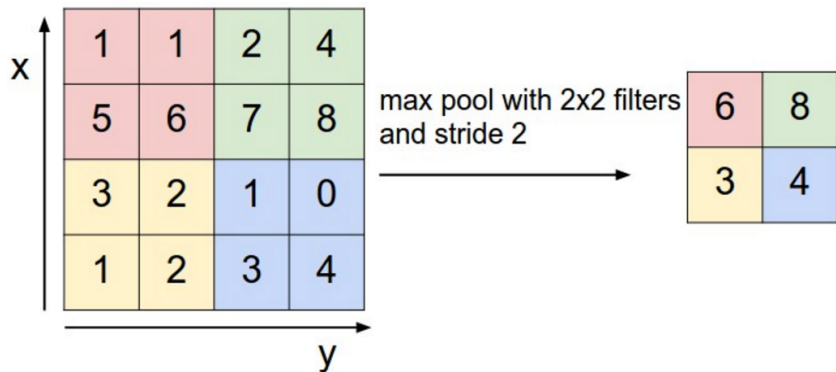
Max-pooling

- По мере продвижения вглубь сверточной нейронной сети, уточняем информацию на картах признаков. Если признак присутствует на предыдущем слое, полученная карта содержит ключевые пиксели, окруженные нечетким «ореолом», в рамках которого признак может быть определен не до конца.
- Решение - метод - max-объединение (max-pooling). Это - разделение карты признаков на непересекающиеся участки и выделения на этих участках нейронов с максимальной активностью. Max-pooling карты признаков делает распознавание более точным, избавляясь от ненужных «ореолов».

Max-pooling



Max-pooling



Max-pooling

- Помимо уменьшения размера тензора, позволяет добавить инвариантность представлению изображения к небольшим поворотам или сдвигам, ускоряет вычисления и привносит в сеть дополнительную нелинейность.
- MaxPooling выбирает из некоторой области максимальное значение, тем самым акцентирует внимание не на расположении какого-либо объекта на изображении, а на самом факте наличия этого объекта. В итоге получаем информацию о содержании изображения.

Субдискретизирующий слой

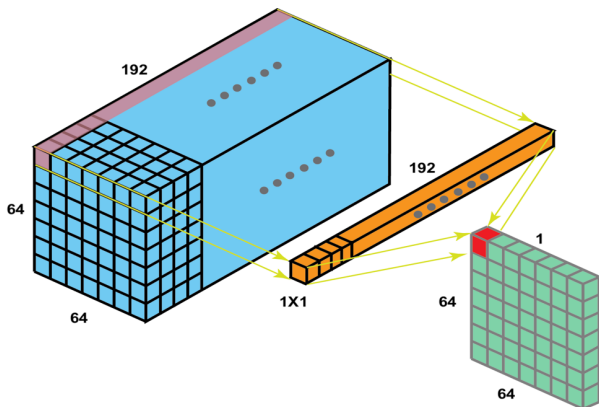
- Выполняет уменьшение размера входной карты признаков, например в 2 раза.
- Разные методы, например, метод выбора максимального элемента (max-pooling) - вся карта признаков разделяется на ячейки 2×2 элемента, из которых выбираются максимальные по значению.
- Формально $\mathbf{x}^{(k)} = f(W^{(k)} \cdot \text{subsample}(\mathbf{x}^{(k-1)}) + b^{(k)})$, f - функция активации, subsample - операция выборки локальных максимальных значений.
- Использование этого слоя позволяет улучшить распознавание примеров с измененным масштабом (уменьшенных или увеличенных).

- В случае монохромного изображения эту сетку реализует двумерный массив размера $H \times W$ (H и W - высота и ширина изображения в пикселях), элементами которого являются значения яркости пикселей из диапазона $[0, 255]$
- Случай цветного изображения: для кодирования цвета могут использоваться различные цветовые схемы, например, XYZ, HSV, или YCbCr, однако самая популярная схема - RGB, в которой цвет пикселя кодируется тремя значениями из диапазона $[0, 255]$, каждое из которых соответствует своему цвету – красному, зеленому или синему. В этом случае изображение - трехмерный тензор размера $H \times W \times C$, где C – количество цветов

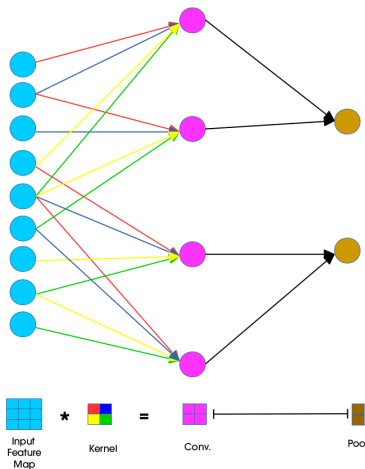
Свчртка с размером ядра 1×1

- Свчртка 1×1 используется для уменьшения количества каналов при введении нелинейности
- При применении свчртки к слишком глубокому тензору (например, $C=256$ - количество цветов), можем предварительно уменьшить его глубину с помощью свчртки 1×1 .
- Т.е. глубину самой сети, при этом на порядки уменьшая количество обучаемых параметров. 1×1 свчртка вместе с некоторой нелинейностью фактически образуют нейронную сеть, принимающую на вход сквозные проекции элементов тензора (“столбцы” размера $1 \times 1 \times C$).

Свёртка с размером ядра 1x1



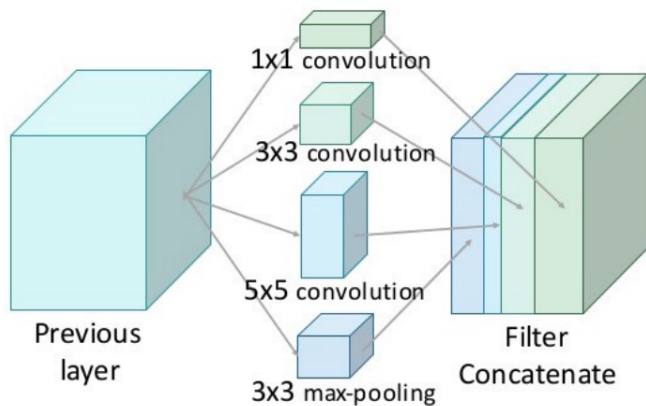
Представление свертки и пулинга



Inception module

- Специальный слой нейронной сети в сети GoogLeNet - конкатенация выходов для сверток размера 1×1 , 3×3 , 5×5 , а также операции max pooling'a с ядром 3×3 .
- Цель:
 - каждый элемент предыдущего слоя соответствует определенной области исходного изображения
 - каждая свертка по таким элементам будет увеличивать область исходного изображения, пока элементы на последних слоях не будут соответствовать всему изображению целиком
 - если с какого-то момента все свертки станут размером 1×1 , то не найдется элементов, которые покрывали бы все исходное изображение, невозможно находить большие признаки

Inception module



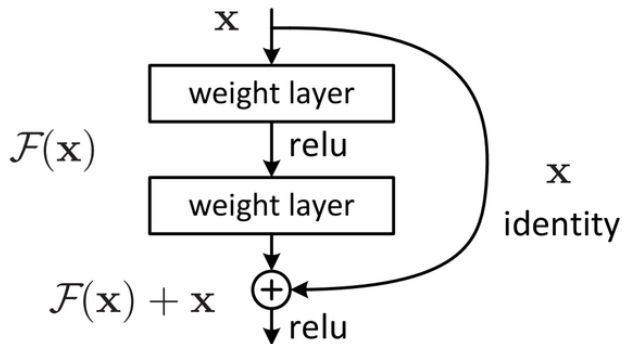
Residual block

- **Проблемы обучения:** исчезающий градиент (vanishing gradient) и взрывающийся градиент (exploding gradient).
- **Причина:** при дифф-нии до глубоких слоев нейронной сети доходит маленькая величина градиента (многократное домножение на малые величины на предыд. слоях).
- **Идея:** residual block! Для пары слоэв (сверточных) добавить доп. связь, которая проходит мимо этих слоэв
 - пусть $z(k)$ - выход k -ого слоя до применения функции активации, а $a(k)$ - выход после.
 - residual block выполняет преобразование:
 $a(k+2) = g(z(k+2) + a(k))$, где g - функция активации.

Residual block

- Сеть обучается предсказывать функцию $F(x) - x$, вместо функции $F(x)$, которую изначально нужно было предсказывать
- Для компенсации этой разницы вводится это замыкающее соединение (shortcut connection), которое добавляет недостающий x к функции.
- Такая разностная функция будет проще обучать, чем исходная.

Residual block (ResNet)



Рецептивные поля (1)

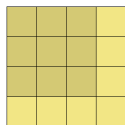
- Рецептивное поле — часть матрицы входных сигналов, подвергаемая “свертыванию”.
- Особенность сверточной сети - размеры ввода становятся все меньше и меньше от начала до конца сети, а количество каналов становится больше
- Receptive field определяет, какую область исходного входа получает выход
- Идея strided convolution - обрабатываем пролеты только на фиксированном расстоянии друг от друга и пропускаем те что посередине
- Оставляем только выходы на определенном расстоянии друг от друга, удаляя остальные

Рецептивные поля (2)

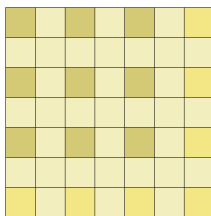
- Процесс свертывания - уплотнит количества данных, заложенных в один пиксел
- Пусть имеется матрица импульсов размерностью 3×3 (9 единиц информации)
- Использование свертки с ядром 2×2 “свернет” входные импульсы в матрицу 2×2 , что даст 4 единицы информации
- Информация в девяти ячейках теперь помещается в меньшем пространстве, в четырех ячейках
- Потери

Рецептивные поля (3)

- Применяем нелинейность к выходным данным, накладываем еще один новый слой свертки сверху.
- Даже если бы применили ядро того же размера (3*3), имеющее одну и ту же локальную область, к выходу strided convolution, ядро имело бы более эффективное receptive field



Operation on the true strided output



Operation on the output with the removed pixels

Рецептивные поля

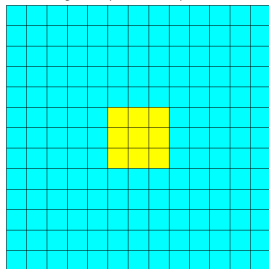
- Это связано с тем, что выход `strided` слоя по-прежнему представляет из себя одно и то же изображение.
- Это не столько обрезка, сколько изменение размера, только теперь каждый отдельный пиксель на выходе является «представителем» большей площади (другие пиксели которой были отброшены) из того же местоположения исходного ввода.
- Поэтому, когда ядро следующего слоя применяется к выходу, оно работает с пикселями, собранными из большей области.
- Такое расширение поля позволяет слоям свертки сочетать признаки низкого уровня (линии, ребра) с признаками более высокого уровня (кривые, текстуры)

Расширяемая свертка (Dilated convolution)

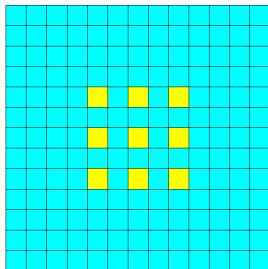
- Расширяемая свертка решает задачу посредством дополнительного слоя нейронной сети
- Похожа на пуллинг и свертку с шагом, но позволяет:
 - расширить рецептивное поле без потери качества изображения
 - получить большее рецептивное поле при тех же затратах на вычисления и расходах памяти, при этом сохранив качество изображения
- Формула свертки: $O_{x,y} = \sum_{i,j} w_{ij} \sum_{|t-i|\leq l, |k-j|\leq l} I_{t,k}$
- I - входные данные, O - выходные, W - ядро свертки, l - коэффициент расширения.

Dilated convolution

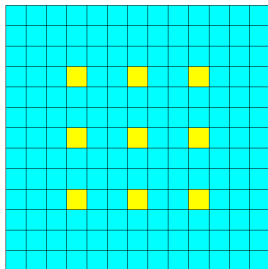
D=1, regular (standard) convol



D=2

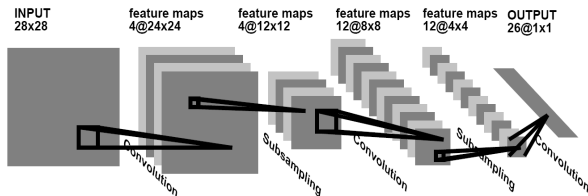


D=3



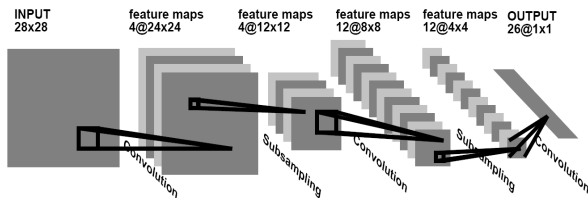
1-, 2- и 4-расширенные свертки с классическими ядрами 3x3, 5x5 и 9x9. Желтые точки обозначают ненулевые веса, остальные веса ядра равны нулю. Выделенные области обозначают рецептивные поля.

Сверточная сеть в целом (1)



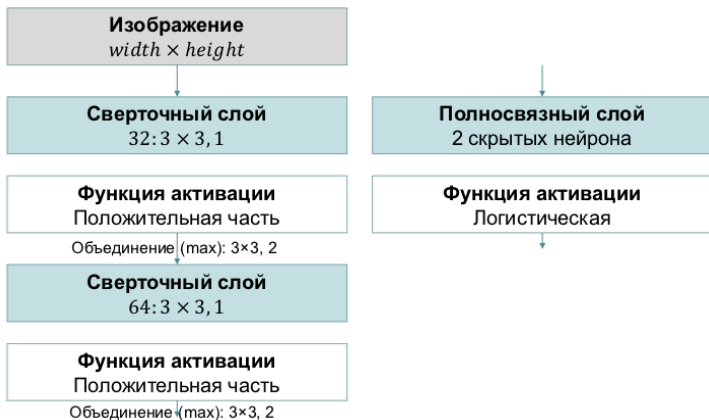
- 1 входной слой - матрица картинки 28×28
- 2 сверточный слой - набор однотипных матриц (карт признаков) по числу ядер свертки
- 3 субдискретизирующий слой - уменьшенный в 2 раза предыдущий набор матриц

Сверточная сеть в целом (2)



- 4 сверточный слой - предыдущий набор матриц суммируется в соответствии со схемой соединения слоев и генерируется новый набор по числу ядер свертки
- 5 субдискретизирующий слой - уменьшенный в 2 раза предыдущий набор матриц
- 6 выходной слой - предыдущий набор матриц разворачивается в вектор и обрабатывается

Пример

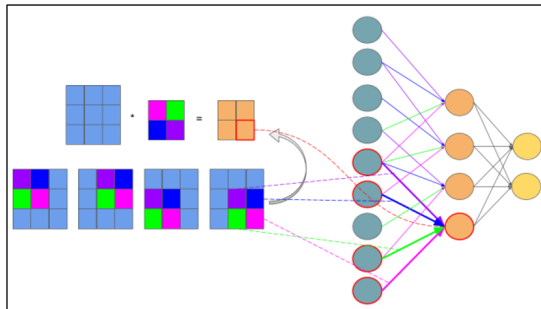


Пример (программа)

```
def generate_cnn_model():
    layers = [
        DataTransform(transform=Normalizer(divisor=128.0)),
        Conv(fshape=(3, 3, 32), padding=2, strides=1,
            dilation=2, init=Kaiming(), activation=Rectlin()),
        Pooling(fshape=(3, 3), padding=1, strides=2, op='max'),
        Conv(fshape=(3, 3, 64), padding=2, strides=1,
            dilation=2, init=Kaiming(), activation=Rectlin()),
        Pooling(fshape=(3, 3), padding=1, strides=2, op='max'),
        Affine(nout=class_count, init=Xavier(),
            activation=Logistic(shortcut=True)) ]
    model = Model(layers=layers)
    cost = GeneralizedCost(costfunc=CrossEntropyBinary())
    return (model, cost)
```

- 1 Случай: сверточный слой находится перед полносвязным, тогда он имеет нейроны и связи такого же типа, как в полносвязном слое, соответственно вычисление δ ошибки ничем не отличается от вычисления δ скрытого слоя.
- 2 Случай: сверточный слой находится перед сверточным, вычисление δ происходит путем обратной свертки.

Обучение сети - случай 1



Интерпретация операции свертки в многослойный вид, где связи с одинаковым цветом имеют один и тот же вес. Синим цветом обозначена подвыборочная карта, разноцветным – синаптическое ядро, оранжевым – получившаяся свертка

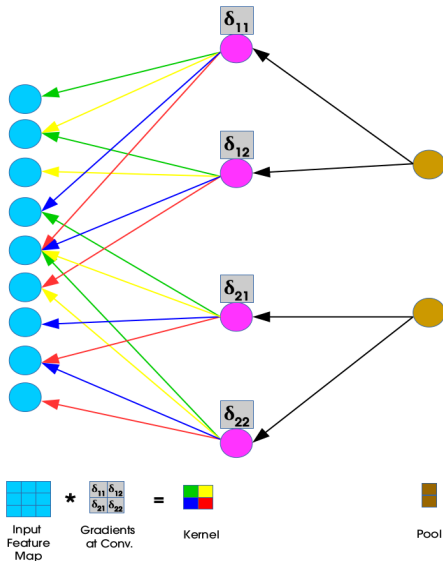
Обучение сети - случай 2

Обратная свертка – это тот же способ вычисления дельт, но различие в повороте ядра на 180 градусов и скользящем процессе сканирования сверточной карты дельт с измененными краевыми эффектами, т.е. необходимо взять ядро сверточной карты (следующего за подвыборочным слоем) повернуть его на 180 градусов и сделать обычную свертку по вычисленным ранее дельтам сверточной карты, но так чтобы окно сканирования выходило за пределы карты.

Обучение сети - случай 2



Обучение сети - случай 2



Особенность соединения слоев и топология сетей (1)

- Сверточную сеть из 7 слоев:
 - 1 входной слой - матрица картинки
 - 2 сверточный слой - набор однотипных карт признаков
 - 3 субдискретизирующий слой - уменьшенный в 2 раза предыдущий набор матриц
 - 4 сверточный слой - предыдущий набор матриц объединяется в одну, в соответствии со схемой соединения слоев, и генерируется новый набор
 - 5 субдискретизирующий слой - уменьшенный в 2 раза предыдущий набор матриц
 - 6 слой MLP - предыдущий набор матриц разворачивается в вектор и обрабатывается как MLP (многослойный перцептрон)
 - 7 слой MLP (выходной)

Особенность соединения слоев (2)

При этом, нейроны (карты признаков) второго субдискретизирующего слоя и третьего сверточного слоя соединяются выборочно в соответствии с матрицей смежности, которая задается как параметр сети. Для сети с количеством карт признаков во втором слое 7 и 9 в третьем слое:

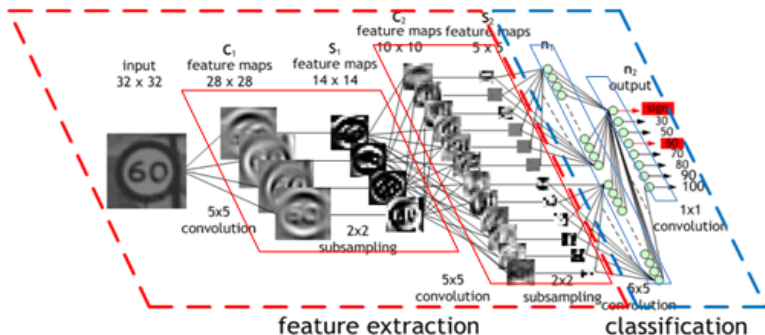
	1	2	3	4	5	6	7	8	9
1	1	0	0	0	0	1	1	1	0
2	1	1	0	0	0	0	1	1	1
3	1	1	1	0	0	0	0	1	1
4	0	1	1	1	0	0	0	1	1
5	0	0	1	1	1	0	0	0	1
6	0	0	0	1	1	1	0	0	0
7	0	0	0	0	1	1	1	0	0

Особенность соединения слоев (3)

- Каждая выходная карта формируется частичной суммой результатов сверток входных карт, для каждой такой частичной суммы свой набор ядер свертки.
- Соединение всех карт второго слоя со всеми картами третьего слоя значительно увеличило бы количество связей.
- Соединение карт одна к одной стало бы еще одним повторением свертки, которое уже присутствовало между слоями.

Сверточная сеть

- Фильтрация 4-мя 5×5 сверточными ядрами, создающими 4 карты признаков.
- Max-pooling
- Фильтрация 10-ю 5×5 сверточными ядрами, max-pooling



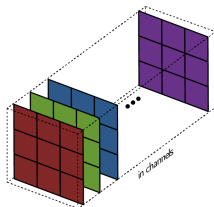
Сверточные сети (формально)

Дано: большие изображения x_{large} размера $r \times c$

- 1 обучаем прореженный автокодер на малых “кусочках” x_{small} размера $a \times b$
- 2 обучаем k признаков $f = \sigma(W^{(1)}x_{\text{small}} + b^{(1)})$ (σ - сигмоид), при наличии весов $W^{(1)}$ и $b^{(1)}$ из входа в нейроны скрытого слоя
- 3 для каждого “кусочка” x_s размера $a \times b$ в большом изображении вычисляем $f_s = \sigma(W^{(1)}x_s + b^{(1)})$
- 4 откуда получаем $f_{\text{convolved}}$ - матрицу “свернутых” признаков размера $k \times (r - a + 1) \times (c - b + 1)$

Многоканальная версия сверточной нейронной сети

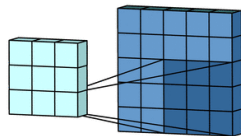
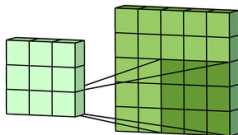
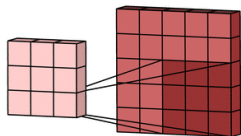
- На практике большинство входных изображений имеют 3 канала (RGB)
- Каждый фильтр - коллекция ядер, причем для каждого отдельного входного канала этого слоя есть одно ядро, и каждое ядро уникально.



Многоканальная версия сверточной нейронной сети

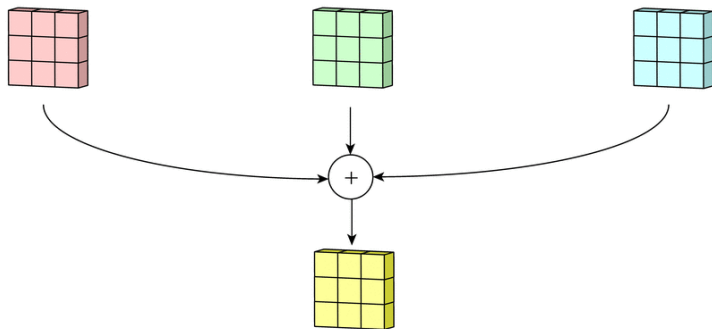
- Каждый фильтр в сверточном слое создает только один выходной канал: каждое из ядер фильтра «скользит» по их соответствующим входным каналам
- Некоторые ядра могут иметь больший вес, чем другие, для того чтобы уделять больше внимания определенным входным каналам (например, фильтр может задать красному каналу ядра больший вес, чем другим каналам).

Многоканальная версия сверточной нейронной сети



Многоканальная версия сверточной нейронной сети

Каждая из обработанных в канале версий суммируется вместе для формирования одного канала.



Преимущества сверточных сетей

- Один из лучших алгоритмов по распознаванию и классификации изображений.
- По сравнению с обычной нейронной сетью гораздо меньшее количество настраиваемых весов, так как одно ядро весов используется целиком для всего изображения, вместо того, чтобы делать для каждого пикселя входного изображения свои весовые коэффициенты.
- Нейросеть при обучении обобщает информацию, а не попиксельно запоминает каждую картинку в весовых коэффициентах, как перцептрон.

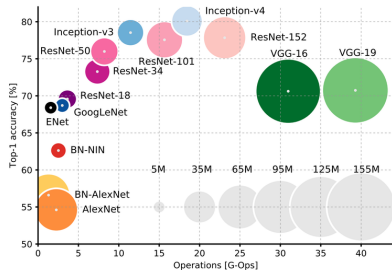
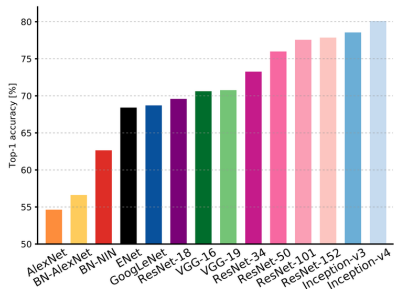
Преимущества сверточных сетей

- Удобное распараллеливание вычислений, возможность реализации на графических процессорах.
- Относительная устойчивость к повороту и сдвигу распознаваемого изображения.
- Обучение при помощи классического метода обратного распространения ошибки.

Недостатки сверточных сетей

1. Архитектура сверточной сети по большей части для распознавания изображений.
2. Слишком много варьируемых параметров сети: количество слоев, размерность ядра свертки для каждого из слоев, количество ядер для каждого из слоев, шаг сдвига ядра при обработке слоя, необходимость слоев субдискретизации, степень уменьшения ими размерности, функция по уменьшению размерности (выбор максимума или среднего) и т.д. Выбираются эмпирически.

Типовые сверточные сети



<http://www.topbots.com/14-design-patterns-improve-convolutional-neural-network-cnn-architecture>

Вопросы

?