CNN Fixations: An unraveling approach to visualize the discriminative image regions

Konda Reddy Mopuri*, Utsav Garg*, R. Venkatesh Babu, *Senior Member, IEEE*

CNN Fixations

"An unraveling approach to visualize the discriminative image regions."

discriminative img regions – special regions which provides special information abt the obj.

Convolutional Neural Networks

2012 – AlexNet, 8 layers, 60 m parameters

↓

2015 – ResNet, 100s of layers, 1,7m parameters

Disadvantage: CNN is a black box.

data → [?] → result

To understand → look at the important img regions that influence their predictions

 → bird

① Вступление

1) Основная идея:

Авторы предлагают способ визуализации того, как CNN принимает решения. Они используют информацию о взаимосвязях м/у выходами нейронов на соседних слоях сети. Когда CNN делает предсказание, активируются опр. нейроны. Это позволяет понять, какие активации на предыдущих слоях сети привели к активации на текущем слое. Такие действия происходят на каждом уровне сети, от softmax (вых. слой) до входного изображения.

Процесс работы:

- выбирается нейрон на каком-либо слое
- Система находит активации на предыдущих слоях, которые поддерживают активацию выбранного нейрона
- повторяется процесс до момента, когда дойдём до вх. изображения
- Т.о., метод помогает определить конкретные участки изображения, которые явл. ответственными за предсказание модели.

Применение:

- метод наглядно даёт понять, например, какие участки изображения позволяют понять, на изображении кошка или собака.
- генерация подписей для изображения.

Преимущества метода:

- модель делается более понятной и прозрачной. Визуализация помогает понять, почему сеть приняла такое решение.
- метод можно использовать не только для финальных активаций, но и для любых др. нейронов на др. уровнях.
- точная покадровая локализация объектов на изображении.

② Сопоставляющая работа.

- Бо́льшая часть работ – градиентный подход: находят области изображения, которые могут улучшить прогнозируемую оценку для выбранной категории.
- Ход и др: карты активации могут быть получены путём объединения карт признаков перед слоем GAP (Global Average Pooling) в соответствии с весами, соединяющими слой GAP с активацией класса в слое классификации.
- Подход, основанный на оценке того, как изм. прогноз, если ф-ция отсутствует.
- → • В отличие от др. работ, подход в статье находит "ответственные" местоположения пикселей, просто распространяя базовые операции прямого прохода через сеть.

Пояснения по работе метода:

Работа начинается с нейрона, который отвечает за предсказанную категорию (например, "кошка"). Далее определяются нейроны, которые активировались на предыдущих слоях. Этот процесс называется "разворачивание" активации (unraveling).

Процесс вывода результата:

Вместо того, чтобы восстанавливать активации (как это делают др. методы), их метод выдаёт бинарный результат на каждом слое сети => становится понятно, какие нейроны были нужны, а какие нет. После этого создаётся тепловая карта (heat map) при помощи размытия бинарного результата Гауссовским размытием (мат. фильтр, который сглаживает картинку).

Простота метода:

Метод не требует настройки параметров (гиперпарам-ов) или др. сложных эвристик (правил).

① Авторы: L.K. Hansen, E.A. Hendricks, N.A. Lydersen, A. Blanchard

Учреждение: Technical University of Denmark, Image and Signal Processing Group

② Год: 2016 year

③ Название сети: CNN (Convolutional Neural Network) (что)

④ Назначение сети: (зачем-задача, входные и выходные данные) Сеть CNN используется для классификации изображений и распознования объектов. Цель состоит в том, чтобы визуализировать дискриминационные области изображения (фиксации), которые CNN использует для принятия решений.

Входные данные: набор изображений, которые используются для обучения и тестирования CNN.

Изображения проходят через несколько слоёв сети (свёрточные слои, пулинг и полносвязные слои), где на каждом уровне происходит обработка и извлечение признаков.

Выходные данные: предсказания сети, то есть метки классов изображений, а также визуальные карты фиксаций, показывающие, какие области изображения были наиболее значимы для принятия решений сети.

Дополнительно с помощью предложенного метода "fixation mapping" выводятся карты фиксаций.

(5) Мотивация авторов к разработке (почему): Мотивация заключается в том, чтобы лучше понять внутр. механизмы работы CNN, которые являются "чёрными ящиками". Несмотря на высокую точность CNN при распознавании объектов и классификации изображений, важно раскрыть, какие части изобр-ий опр. выводы сети.

Это помогает:

- улучшить доверие к работе НС, особенно в ответственных приложениях, таких как мед. и беспилотные системы.
- опр. слабые места сети, которые можно улучшить для более точной классификации.
- ускорение разработки новых арх. CNN

⑥ Состав и отличие от базовой архитектуры (как решают):

В осн. ленит стандартная структура CNN, но авторы добавили метод визуализации "fixation mapping". Этот метод позволяет на каждом слое сети выделить те области изображения, которые оказывают наиб. влияние на предсказания сети.

*Важная особенность: не требует изм. в базовой архитектуре CNN, но существенно расширяет понимание того, как сеть "видит" из-я на разных уровнях. В отличие, например, от метода Grad-CAM, фиксации предлагают более прямую визуализацию ключевых областей.

- Grad-CAM – Gradient-weighted Class Activation Method.
- фиксации – показывает ключевые обл. изображения.
- активации – визуализирует отклик нейронов на изображение.

⑦ Качественные и количественные показатели: (как оценивают или сравнивают)

1) Качественные показатели: карты фиксаций
2) Количественные —"—:
- точность классификации CNN, когда сети даются только зафиксированные участки из-я.
- происходит сравнение с исх. изображением и др. методами визуализации.

⑧ Есть ли потенциал развития (плюсы/минусы):

⊕: 1) Возможность улучшения интерпретируемости НС, что важно для таких приложений с использованием НС, например для классификаций изображений или в мед. отрасли для анализа снимков.
2) Простота интеграции с сущ. арх-ми CNN.
3) Возможность исп-я метода для улучшения обучения НС, направляя внимание на ключ. эл-ты изображений.

⊖: 1) Сложно обобщить для более сложных и разнообр. наборов данных
2) Ограничение на типы сетей. Метод лучше всего подходит для CNN.

Как работают сверточные нейронные сети (CNN)?

Сверточные нейронные сети отличаются от других нейронных сетей превосходной производительностью при вводе изображений, речи или аудиосигнала. Они имеют три основных типа слоев, которые бывают:

- Сверточный слой
- Слой пулинга
- Полносвязный (FC) слой

Сверточный слой — это первый слой сверточной сети. В то время как за сверточными слоями могут следовать дополнительные сверточные слои или объединяющие слои, полностью соединенный слой является последним. С каждым слоем CNC становится все сложнее, идентифицируя все большие части изображения. Более ранние слои сосредоточены на простых элементах, таких как цвета и края. По мере того, как данные изображения проходят через слои CNC, она начинает распознавать более крупные элементы или формы объекта, пока окончательно не идентифицирует предполагаемый объект.

1. Сверточный слой

Сверточный слой является основным строительным блоком CNC, и именно на нем происходит большая часть вычислений. Для этого требуется несколько компонентов: входные данные, фильтр и карта функций. Предположим, что входными данными будет цветное изображение, которое в 3D состоит из матрицы пикселей. Это означает, что входные данные будут иметь три измерения — высоту, ширину и глубину, — которые соответствуют RGB в изображении. У нас также есть детектор признаков, также известный как ядро или фильтр, который будет перемещаться по рецептивным полям изображения, проверяя, присутствует ли признак. Этот процесс известен как свертка.

Детектор признаков представляет собой двумерный (2D) массив весов, который представляет собой часть изображения. Хотя они могут различаться по размеру, размер фильтра обычно представляет собой матрицу 3x3; Это также определяет размер рецептивного поля. Затем фильтр применяется к области изображения, а между входными пикселями и фильтром вычисляется точечное произведение. Затем это скалярное произведение подается в выходной массив. После этого фильтр быстро смещается, повторяя процесс до тех пор, пока ядро не пройдет по всему изображению. Окончательные выходные данные ряда точечных произведений входных

данных и фильтра называются картой признаков, картой активации или сверточным объектом.

Обратите внимание, что веса в детекторе признаков остаются неизменными при его перемещении по изображению, что также известно как совместное использование параметров. Некоторые параметры, такие как значения веса, корректируются во время тренировки в процессе обратного распространения и градиентного спуска. Однако есть три гиперпараметра, влияющих на размер объема выходных данных и которые необходимо задать до начала обучения нейронной сети. К ним относятся:

1. **Количество фильтров** влияет на глубину вывода. Например, три различных фильтра дадут три разные карты признаков, создав глубину в три.

2. **Шаг** — это расстояние, или количество пикселей, на которое ядро перемещается по входной матрице. В то время как значения шага два или больше встречаются редко, больший шаг дает меньший результат.

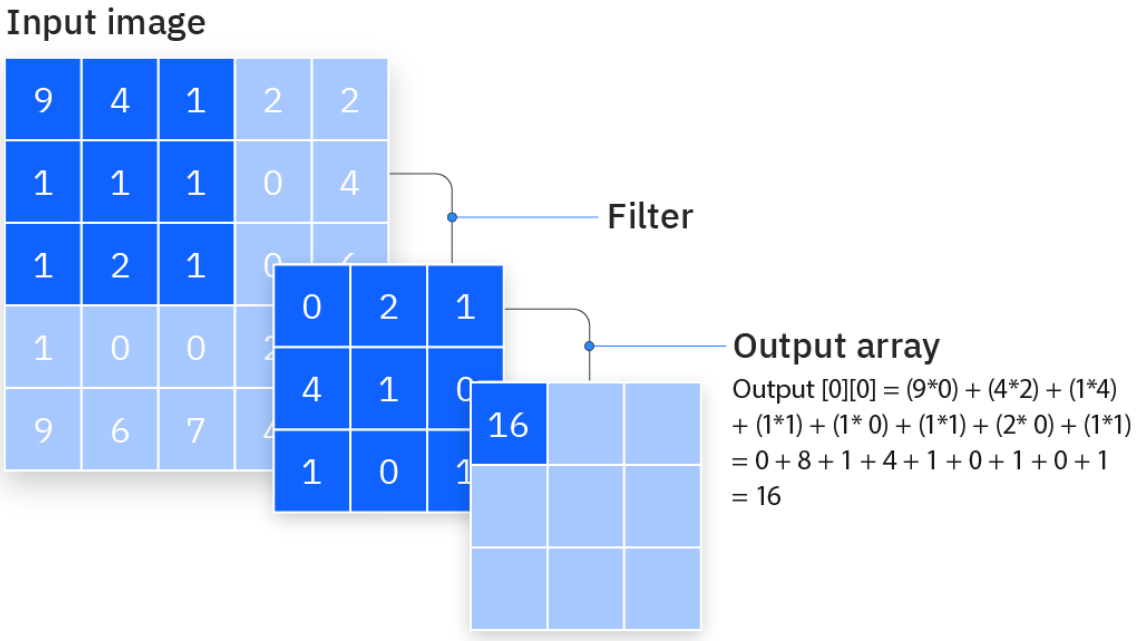
3. **Нулевое отступление** обычно используется, когда фильтры не подходят к входному изображению. При этом все элементы, находящиеся за пределами входной матрицы, обнуляются, что приводит к большему или равному по размеру выходу. Существует три вида набивки:

Допустимые отступы: Это также известно как отсутствие набивки. В этом случае последняя свертка отбрасывается, если размеры не совпадают.

Такая же набивка: Это заполнение гарантирует, что выходной слой будет иметь тот же размер, что и входной слой.

Полная набивка: Этот тип заполнения увеличивает размер вывода за счет добавления нулей к границе входа.

После каждой операции свертки CHC применяет преобразование выпрямленных линейных единиц (ReLU) к карте признаков, внося в модель нелинейность.



2. Дополнительный сверточный слой

Как мы упоминали ранее, за начальным слоем свёртки может следовать другой слой свертки. Когда это происходит, структура CHC может стать иерархической, поскольку более поздние слои могут видеть пиксели в пределах восприимчивых полей предыдущих слоев. В качестве примера предположим, что мы пытаемся определить, есть ли на изображении велосипед. Велосипед можно рассматривать как сумму частей. Он состоит из рамы, руля, колес, педалей и так далее. Каждая отдельная часть велосипеда составляет низкоуровневый шаблон в нейронной сети, а комбинация его частей представляет собой более высокоуровневый шаблон, создавая иерархию признаков внутри CHC. В конечном счете, сверточный слой преобразует изображение в числовые значения, позволяя нейронной сети интерпретировать и извлекать соответствующие закономерности.

3. Слой пулинга

Объединение слоев, также известное как даунсэмплинг, приводит к уменьшению размерности, уменьшая количество параметров на входе. Как и в случае со сверточным слоем, операция объединения охватывает фильтр по всему входу, но разница заключается в том, что этот фильтр не имеет весовых коэффициентов. Вместо этого ядро применяет функцию агрегирования к значениям в восприимчивом поле, заполняя выходной массив. Существует два основных типа пулинга:

- **Максимальный пул:** По мере того, как фильтр перемещается по входным данным, он выбирает пиксель с максимальным значением для отправки в выходной массив. Кстати, этот подход, как правило, используется чаще по сравнению со средним пулингом.
- **Средний пул:** Когда фильтр перемещается по входным данным, он вычисляет среднее значение в восприимчивом поле для отправки в выходной массив.

Несмотря на то, что на уровне пула теряется много информации, это также имеет ряд преимуществ для CNN. Они помогают снизить сложность, повысить эффективность и ограничить риск переобучения.

4. Полносвязный слой

Название полносвязного слоя точно описывает само себя. Как упоминалось ранее, значения пикселей входного изображения не связаны напрямую с выходным слоем в частично соединенных слоях. Однако в полносвязном слое каждый узел в выходном слое подключается непосредственно к узлу в предыдущем слое.

Этот слой выполняет задачу классификации на основе признаков, извлеченных через предыдущие слои, и их различных фильтров. В то время как сверточные слои и слои пула, как правило, используют функции ReLu, уровни FC обычно используют функцию активации softmax для соответствующей классификации входных данных, получая вероятность от 0 до 1.

Типы сверточных нейронных сетей

Кунихико Фукусима и Ян Лекун заложили основу исследований сверточных нейронных сетей в своей работе в 1980 году (ссылка находится за пределами ibm.com) и «Обратное распространение применительно к распознаванию рукописного почтового индекса» в 1989 году соответственно. Более известен тот факт, что Ян Лекун успешно применил обратное распространение ошибки для обучения нейронных сетей выявлению и распознаванию закономерностей в серии рукописных почтовых индексов. Он продолжил свои исследования со своей командой в течение 1990-х годов, достигнув кульминации в «LeNet-5», который применил те же принципы предыдущих исследований к распознаванию документов. С тех пор появилось несколько вариантов архитектур CNN с введением новых наборов данных, таких как MNIST и CIFAR-10, а также конкурсов, таких как ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Некоторые из этих других архитектур включают в себя:

- AlexNet (ссылка находится за пределами ibm.com)
- VGGNet (ссылка находится за пределами ibm.com)
- GoogLeNet (ссылка находится за пределами ibm.com)
- ResNet (ссылка находится за пределами ibm.com)
- ZFNet

  Тем не менее, LeNet-5 известен как классическая архитектура CNN.

Сверточные нейронные сети и компьютерное зрение

Сверточные нейронные сети обеспечивают распознавание изображений и задачи компьютерного зрения. Компьютерное зрение — это область искусственного интеллекта (ИИ), которая позволяет компьютерам и системам извлекать значимую информацию из цифровых изображений, видео и других визуальных данных, и на основе этих входных данных они могут принимать меры. Эта способность давать рекомендации отличает его от задач распознавания изображений. Некоторые распространенные применения этого компьютерного зрения сегодня можно увидеть в:

- **Маркетинг:** Платформы социальных сетей предоставляют предложения о том, кто может быть на фотографии, опубликованной в профиле, что упрощает отметку друзей в фотоальбомах.
- **Здравоохранение:** Компьютерное зрение было включено в радиологические технологии, что позволяет врачам лучше выявлять раковые опухоли в здоровой анатомии.
- **Розничный:** Визуальный поиск был встроен в некоторые платформы электронной коммерции, что позволяет брендам рекомендовать вещи, которые дополнят существующий гардероб.
- **Автомобильная промышленность**: В то время как эпоха беспилотных автомобилей еще не совсем наступила, лежащая в ее основе технология начала проникать в автомобили, повышая безопасность водителя и пассажиров с помощью таких функций, как определение линии полосы движения.