УДК 004.89

ПРИМЕНЕНИЕ ВИЗУАЛЬНЫХ СРЕДСТВ ДЛЯ ИНТЕРПРЕТАЦИИ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ

Репьев И.В., студент гр. 24.М02-мкн, I курс
Санкт-Петербургский государственный университет, г. Санкт-Петербург

Сверточные нейронные сети (CNN, convolutional neural networks) стали краеугольным камнем современной компьютерной зрительной аналитики, предоставив прорывные результаты в распознавании образов, детекции объектов, сегментации сцен и многом другом. Однако, несмотря на высокую точность таких моделей, они часто рассматриваются как «чёрные ящики» из-за своей многослойной иерархической структуры, внутри которой скрыто огромное количество параметров и фильтров, чьё влияние на итоговое решение трудно объяснить напрямую.

В условиях, когда модели CNN применяются в чувствительных областях – медицинская диагностика по изображениям, автономные транспортные системы, системы безопасности – растёт необходимость не только в точных, но и в объяснимых предсказаниях, способных быть интерпретированными как экспертами, так и конечными пользователями. В этом контексте визуальные методы интерпретации CNN становятся не просто инструментом анализа, но частью инфраструктуры доверенного и подотчётного искусственного интеллекта.

Основная идея визуальной интерпретации заключается в отображении, какие регионы изображения активируют определённые нейроны или слои сети и как эти активации связаны с финальным решением. Один из первых и до сих пор широко используемых методов – это карты активации (activation maps), которые отображают уровни возбуждения различных фильтров при прохождении изображения через слои сети. Такие карты позволяют исследовать, какие части изображения важны для конкретных фильтров, и как нарастает абстракция признаков от низкоуровневых (грани, текстуры) к высокоуровневым (формы, объекты). Однако такие карты дают лишь косвенное представление о важности признаков для конкретного класса и не позволяют точно локализовать критические зоны, влияющие на итоговое предсказание.

В этой связи наибольшее распространение получили методы Grad-CAM (Gradient-weighted Class Activation Mapping) и его модификации, позволяющие получить тепловую карту, отражающую вклад каждого региона изображения в активацию конкретного выхода (например, принадлежность к определённому классу). Grad-CAM использует градиенты, проходящие через последние

сверточные слои, чтобы взвесить карты признаков и агрегировать их в виде наглядной визуализации.

Результирующая карта наложения (heatmap) показывает, какие участки изображения были наиболее значимыми для принятия решения моделью. Этот метод эффективен тем, что позволяет локализовать внимание модели без необходимости модификации архитектуры или переобучения, и может применяться к уже готовым предобученным моделям.

Продолжением Grad-CAM стал Score-CAM, устраняющий зависимость от градиентов и использующий непосредственно активации слоёв и влияние маскированных входов на выход. Этот метод более устойчив к шумам и хорошо работает на сложных изображениях, где градиентная информация становится слишком размытой. В медицинской визуализации, например, при диагностике пневмонии по рентгеновским снимкам, Score-CAM может более точно локализовать патологические области, чем обычный Grad-CAM, тем самым повышая доверие к выводу модели у врача.

Другим направлением является визуализация фильтров и нейронных предпочтений. Методы типа feature visualization или activation maximization создают изображения, которые максимально активируют конкретный нейрон или фильтр. Такие «портреты» признаков могут быть представлены как формы, текстуры или даже объектоподобные структуры, на которые «настроен» конкретный канал. Это позволяет исследовать специализацию различных фильтров и строить интуитивные карты «ответственности» слоёв за разные визуальные паттерны. В обучающих целях такие методы играют важную роль в объяснении архитектурных решений и позволяют разработчикам понимать, где в сети возникают проблемы, связанные с переобучением или недостаточной обобщающей способностью.

Отдельное внимание уделяется визуализации на уровне входного изображения, где используется метод saliency maps – отображения градиентов предсказания по отношению к каждому пикселю входа. Это позволяет определить, как чувствительно предсказание модели к изменению конкретных областей изображения. Хотя saliency карты имеют тенденцию к высокой разреженности и шумности, они предоставляют высокую пространственную точность и могут служить основой для построения контрфактических примеров или тестирования устойчивости модели к локальным искажениям. Более стабильные модификации, такие как SmoothGrad, Integrated Gradients или Guided Backpropagation, позволяют сгладить карту важности и сделать визуализацию более интерпретируемой человеком.

Интерес представляет и совмещение визуальных интерпретаций с семантической сегментацией, где активации и карты внимания накладываются на сегментированные объекты или анатомические структуры. Такой подход

полезен, например, в анализе медицинских изображений, когда необходимо объяснить, что модель приняла решение на основе конкретного участка ткани или органа, а не фона или артефактов. Такие методы сочетают визуализацию внимания модели с внешними онтологиями и формируют интерпретации, согласующиеся с экспертными знаниями.

Современные платформы и библиотеки, такие как Captum (для PyTorch), TF-Explain (для TensorFlow), Lucid, Alibi Explain и Netron, предоставляют широкий инструментарий для генерации и оценки визуальных объяснений. Эти средства позволяют не только строить тепловые карты и визуализировать фильтры, но и логгировать объяснения, сравнивать поведение модели на разных входах, а также генерировать визуальные отчёты для аудита и документации. Всё чаще такие визуализации становятся неотъемлемой частью интерфейса работы с моделью – особенно в тех приложениях, где человек должен совместно с ИИ принимать решение, и объяснение должно быть интуитивно понятным и пространственно локализованным.

Визуальные интерпретации CNN применяются также для оценки доверия и устойчивости модели. Например, сравнение карт активации при исходном изображении и при изображении с добавленным шумом или незначительной трансформацией позволяет выявить, насколько стабильно поведение модели. Если тепловая карта «прыгает» между различными участками изображения при минимальных изменениях входа, это может указывать на переобучение или уязвимость к атакующим шумам. Аналогично, сравнение карт внимания между различными классами позволяет диагностировать наличие предвзятости – например, если модель уделяет внимание этническим или визуальным признакам, не имеющим отношения к целевой задаче.

Таким образом, визуальные средства интерпретации сверточных нейронных сетей играют ключевую роль в обеспечении прозрачности, подотчётности и доверия к компьютерному зрению [1].

Они позволяют сделать поведение сложных моделей наглядным, объяснимым и доступным как техническим специалистам, так и конечным пользователям [2].

Более того, визуальные интерпретации становятся инструментом не только понимания модели, но и её верификации, тестирования, отладки и документирования, тем самым занимая центральное место в инженерии интерпретируемого искусственного интеллекта [3-6].

Список литературы:

1. Zhou, B. Learning Deep Features for Discriminative Localization / B. Zhou, A. Khosla, A. Lapedriza [et al.] // 2016 IEEE Conference on Computer Vision and

Pattern Recognition. – Las Vegas : IEEE, 2016. – P. 2921–2929. – DOI 10.1109/CVPR.2016.319.

2. Chattopadhay, A. Grad-CAM++ : Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks / A. Chattopadhay, A. Sarkar, P. Howlader, V. N. Balasubramanian // 2018 IEEE Winter Conference on Applications of Computer Vision. – Lake Tahoe : IEEE, 2018. – P. 839–847. – DOI 10.1109/WACV.2018.00097.

3. Mahendran, A. Visualizing Deep Convolutional Neural Networks Using Natural Pre-Images / A. Mahendran, A. Vedaldi // International Journal of Computer Vision. – 2016. – Vol. 120. – P. 233–255. – DOI 10.1007/s11263-016-0911-8.

4. Olah, C. Feature Visualization / C. Olah, A. Mordvintsev, L. Schubert // Distill. – 2017. – DOI 10.23915/distill.00007.

5. Carter, S. Activation Atlas / S. Carter, Z. Armstrong, L. Schubert [et al.] // Distill. – 2019. – DOI 10.23915/distill.00015.

6. Свидетельство о государственной регистрации программы для ЭВМ № 2024616661 Российская Федерация. Программа для выполнения автоматизированной обработки изображений и видео на основе компьютерного зрения : № 2024614663 : заявл. 09.03.2024 : опубл. 22.03.2024 / П. А. Пылов. – EDN TXKINN.