Методы оптимизации нейронных сетей для анализа многоспектральных данных

Н. А. Харковчук, Е. Ю. Авксентьева, И. С. Мухин

Университет ИТМО

nakharkovchuk@itmo.ru

Аннотация. **В работе исследуются методы оптимизации сверточных (CNN) сетей различных вариантов автокодировщиков для анализа многоспектральных данных, в том числе, для работы в условиях ограниченных вычислительных ресурсов. Рассмотрены различные архитектуры нейронных сетей, включая гибридные модели, сочетающие пространственную экстракцию признаков с моделированием спектральных зависимостей, а также методы оптимизации: квантование весов и активаций (int8), кластеризацию и аппаратно-зависимые ускорения. Применение полного целочисленного квантования позволяет сократить время инференса порядка на 65% при минимальной деградации точности около 5%, обеспечивая устойчивую частоту обработки кадров выше 25 FPS на целевой платформе. Полученные результаты демонстрируют эффективность предложенных подходов для развертывания ресурсоемких моделей на периферийных устройствах с сохранением высокой точности для потоковой классификации многоспектральных данных.**

Ключевые слова: многоспектральные данные; автоэнкодер; сверточные нейронные сети; оптимизация моделей; прунинг; квантование; ONNX

I. Обзор литературы

Современные методы дистанционного зондирования Земли обеспечивают получение мульти- и гиперспектральных изображений, содержащих десятки и сотни спектральных каналов. Как отмечают авторы в [1], ключевая концепция многоспектрального зондирования заключается в том, что каждый пиксель изображения характеризуется уникальным спектральным откликом, зависящим от типа подстилающей поверхности. Однако высокая размерность данных, наличие шумов и зависимость от условий съёмки существенно затрудняют их интерпретацию.

В последние годы для решения задач анализа спектральных данных активно применяются методы глубокого обучения [2]. Особое место занимают свёрточные нейронные сети (CNN), которые, как показано в работах [3, 4], эффективны для извлечения иерархических признаков. Применительно к гиперспектральным данным авторы в [5] предложили использование сложенного автоэнкодера (SAE) для формирования глубоких признаков. Архитектура однослойного автоэнкодера, включающая энкодер и декодер с привязанными весами, описана в [6]; обучение при этом направлено на минимизацию ошибки реконструкции.

Для классификации гиперспектральных изображений Hu et al. [7] разработали модель REFEREE, основанную на рекуррентных нейронных сетях (RNN) и предназначенную для обнаружения изменений по разновременным снимкам. Разработчики используют блоки LSTM для кодирования мультивременного входа. Дальнейшее развитие этот подход получил в гибридных архитектурах CRNN [8, 9], где свёрточные слои извлекают локальные пространственные признаки, а рекуррентные слои моделируют зависимости между областями изображения.

Вместе с тем проведённый анализ научных решений выявляет ряд систематических ограничений. Во многих работах отсутствует информация о спектральных диапазонах, для которых производилась обработка; не применяются архитектуры на основе трансформеров; не производится разделение по регионам и типам съёмочных платформ (спутник, БПЛА). Кроме того, большинство исследований сосредоточено либо исключительно на предварительной обработке и обучении конкретной архитектуры, либо на достижении максимальной точности без учёта вычислительных затрат. Практически не рассматриваются вопросы оптимизации моделей для работы на периферийных устройствах с ограниченными ресурсами памяти и производительности, а также сравнительный анализ эффективности автоэнкодерных и свёрточных архитектур при решении задач реконструкции и регрессии на многоспектральных данных.

Таким образом, несмотря на доказанную эффективность нейросетевых подходов, остаётся открытым вопрос о том, какие архитектуры (автоэнкодеры или свёрточные сети) и какие методы оптимизации (прунинг, квантование, преобразование в ONNX) обеспечивают наилучший компромисс между точностью и вычислительной эффективностью при обработке многоспектральных данных, в условиях ограниченных ресурсов. Настоящая работа направлена на восполнение этого пробела путём сравнительного исследования указанных архитектур и методов оптимизации применительно к задачам реконструкции и регрессионного анализа спектральных данных.

II. Основная часть

В рамках проведённого исследования для каждой из рассматриваемых архитектур была сформирована совокупность экспериментальных конфигураций, отличающихся степенью структурной и численной оптимизации модели. Такой подход позволил оценить влияние различных методов снижения вычислительной сложности на качество работы моделей. Базовая конфигурация модели использовалась в качестве эталонной и не включала дополнительных процедур оптимизации.

Базовая конфигурация отражает исходные характеристики архитектуры и служит отправной точкой для сравнительного анализа. Конфигурация с применением прунинга предполагала удаление части весовых коэффициентов, оказывающих наименьшее влияние на выход модели. Отбор параметров осуществлялся на основе их вклада в формирование сигналов функций активаций. Применение данного метода направлено на снижение избыточности модели и уменьшение её размерности без существенного изменения функционального поведения. Конфигурация с квантованием включала преобразование весов и функций активаций для работы с пониженной разрядностью. Данный подход позволяет уменьшить объём занимаемой памяти и ускорить вычисления, особенно при выполнении на центральных процессорах и встраиваемых системах. При этом исследуется влияние снижения точности представления параметров на итоговые метрики качества. Отдельная конфигурация предполагала представление модели в стандартизированном виде вычислительного графа. Использование данного формата обеспечивает более эффективное выполнение модели за счёт оптимизации вычислений на уровне среды исполнения и возможности применения аппаратно-зависимых ускорений.

Для автоэнкодерной архитектуры дополнительно была рассмотрена модификация, связанная с изменением размерности латентного пространства. В рамках данной конфигурации варьировалось число каналов и степень пространственного сжатия признакового представления. Это позволило оценить влияние степени компрессии на способность модели к восстановлению исходных данных.

Таким образом, совокупность экспериментальных конфигураций охватывает как структурные изменения архитектуры, так и методы численной оптимизации, что обеспечивает комплексный анализ их влияния на характеристики моделей.

Рассматриваемая автоэнкодерная модель представляет собой сверточную нейронную архитектуру, предназначенную для извлечения компактного представления входных данных с последующим восстановлением исходного изображения. Архитектура построена по принципу композиции двух взаимосвязанных модулей: энкодера и декодера.

Энкодер реализует отображение входного тензора изображения в пространство скрытых представлений пониженной размерности. Структурно он представлен последовательностью сверточных блоков, каждый из которых включает операцию свертки, нормализацию и нелинейную активацию. На каждом уровне энкодера происходит уменьшение пространственного разрешения признакового представления за счёт использования сверток с шагом больше единицы либо операций подвыборки. Одновременно с этим увеличивается число каналов, что позволяет компенсировать потерю пространственной информации за счёт расширения признакового пространства. В отличие от классических автоэнкодеров, в данной архитектуре не осуществляется полное преобразование входного изображения в одномерный вектор фиксированной длины. Латентное представление сохраняет двумерную структуру, что обеспечивает сохранение локальных пространственных зависимостей между элементами изображения. Такая особенность является принципиально важной при обработке многоспектральных данных, где пространственная согласованность каналов играет существенную роль.

Сформированное латентное пространство является результатом работы энкодера и представляет собой компактное кодирование исходного изображения с пониженной пространственной размерностью и увеличенным числом каналов. Сохранение двумерной структуры латентного пространства позволяет учитывать локальные закономерности в данных, что особенно важно для задач реконструкции изображений. Размерность латентного пространства определяется компромиссом между степенью сжатия и точностью восстановления.

Декодер реализует отображение из латентного пространства обратно в пространство исходных изображений. Он состоит из последовательности блоков, выполняющих увеличение пространственного разрешения и уменьшение числа каналов. Для восстановления размерности используются транспонированные сверточные операции, а также операции интерполяции. На каждом уровне декодера происходит постепенное восстановление пространственной структуры изображения. Симметричность архитектуры декодера относительно энкодера способствует более устойчивому процессу обучения и улучшает качество реконструкции. Ключевой особенностью рассматриваемой архитектуры является сохранение пространственной структуры латентного представления, что отличает её от классических автоэнкодеров с полносвязными слоями и позволяет более эффективно работать с изображениями. Дополнительно, отказ от чрезмерного сжатия данных способствует снижению потерь информации, что положительно сказывается на качестве реконструкции. Увеличение числа каналов в латентном пространстве компенсирует уменьшение пространственного разрешения. Таким образом, архитектура автоэнкодера обеспечивает баланс между степенью компрессии данных и сохранением информативных признаков.

A. Сверточная нейронная сеть

Сверточная нейронная сеть, рассматриваемая в работе, реализует архитектуру с общим блоком извлечения признаков и несколькими выходными ветвями, предназначенными для решения различных задач. Такой подход соответствует парадигме многозадачного обучения и позволяет эффективно использовать общее представление данных.

Базовый модуль извлечения признаков осуществляет начальный этап обработки данных с помощью последовательности сверточных блоков, формирующих иерархическое представление признаков. Каждый блок включает: операцию свертки, обеспечивающую локальную агрегацию информации; нормализацию, способствующую стабилизации процесса обучения; нелинейную активацию, операцию подвыборки, уменьшающую пространственную размерность. В процессе прохождения через данные блоки формируется признаковое пространство высокой размерности, в котором закодированы как локальные, так и глобальные характеристики входных данных.

Агрегация признаков применяется после прохождения сверточных слоёв и позволяет преобразовать тензор признаков в компактное представление. В качестве такой операции может использоваться глобальное усреднение по пространственным координатам. Данный этап обеспечивает переход от пространственного представления к векторному, пригодному для последующей обработки полносвязными слоями.

Выходные ветви представлены двумя независимыми компонентами. Классификационная ветвь предназначена для решения задачи классификации и включает один или несколько полносвязных слоёв, завершающихся функцией активации, соответствующей типу задачи. При многоклассовой классификации используется сигмоидальная активация, позволяющая независимо оценивать вероятность принадлежности к каждому классу. Регрессионная ветвь используется для предсказания непрерывных значений и состоит из полносвязных слоёв с линейной активацией на выходе, обеспечивая оценку количественных характеристик, связанных с входными данными.

B. Ветви обучения и функциональные особенности

Использование общей сверточной части и нескольких выходных ветвей позволяет реализовать многозадачное обучение, при котором модель одновременно оптимизирует несколько функций потерь. Такой подход способствует формированию более устойчивого и информативного признакового представления, а совместное обучение различных задач позволяет учитывать взаимосвязи между ними, повышая качество предсказаний по сравнению с обучением отдельных моделей. Ключевым преимуществом архитектуры является возможность одновременного решения задач различной природы на основе единого признакового пространства, что обеспечивает более эффективное использование данных и снижает общую вычислительную сложность. Разделение на специализированные выходные ветви позволяет адаптировать модель к различным типам выходных данных, сохраняя при этом общую структуру и логику обработки информации.

III. Результаты

В рамках проведённого исследования была выполнена комплексная оценка качества и вычислительной эффективности двух классов моделей: сверточных нейронных сетей, решающих задачу регрессии долей классов, и автоэнкодерных архитектур, ориентированных на реконструкцию входных данных. Эксперименты проводились для различных конфигураций входных данных, включая использование только RGB-каналов и расширенных мультимодальных представлений с добавлением спектральных индексов. Дополнительно исследовано влияние методов оптимизации моделей.

Результаты оценки качества сверточных моделей представлены в табл. 1.

ТАБЛИЦА I. Результаты сверточных моделей

Конфигурация модели	**MAE (micro)**	**RMSE (micro)**	**R^2 (macro)**
CNN (RGB)	0.0809	0.1329	0.1118
CNN (RGB + индексы)	0.0729	0.1304	0.1553

Анализ результатов показывает, что использование спектральных индексов приводит к улучшению качества модели. Наблюдается снижение средней абсолютной ошибки и увеличение коэффициента детерминации, что свидетельствует о более точном восстановлении распределения классов.

Данный эффект объясняется тем, что спектральные индексы представляют собой информативные признаки, напрямую связанные с физическими характеристиками наблюдаемой поверхности. В отличие от исходных RGB-каналов, они позволяют выделить структурные особенности растительности и почвы, что облегчает задачу регрессии. Таким образом, в рамках сверточной модели дополнительная информация используется эффективно и приводит к улучшению качества.

Результаты для автоэнкодерных моделей, включая базовый сверточный автоэнкодер и более сложную мультимодальную архитектуру, представлены в табл. 2 и табл. 3.

ТАБЛИЦА II. Результаты автоэнкодеров

Конфигурация модели	**Loss**	**MAE (micro)**	**RMSE (micro)**	**R^2 (micro)**
Autoencoder (RGB)	0.4118	0.0557	0.1197	0.6456
Autoencoder (RGB + индексы)	0.3647	0.0653	0.1208	0.6393

ТАБЛИЦА III. Результаты автоэнкодеров

Конфигурация модели	**Loss**	**MAE (macro)**	**RMSE (macro)**	**R^2 (macro)**
Autoencoder (RGB)	0.4118	0.055774	0.094452	0.2719
Autoencoder (RGB + индексы)	0.3647	0.065365	0.096718	0.2695

Полученные результаты демонстрируют, что добавление спектральных индексов приводит к снижению значения функции потерь, однако сопровождается увеличением средней абсолютной ошибки и незначительным снижением коэффициента детерминации. Это указывает на более сложный характер влияния дополнительных признаков в задаче реконструкции.

Снижение функции потерь свидетельствует о том, что модель лучше аппроксимирует глобальную структуру входных данных. Однако увеличение ошибки указывает на ухудшение локальной точности восстановления. Это может быть связано с тем, что спектральные индексы являются производными от исходных каналов и вводят избыточность, усложняя процесс обучения.

Дополнительные эксперименты с комплексной архитектурой автоэнкодера, основанной на раздельной обработке RGB-данных и спектральных индексов, показали отсутствие существенного прироста качества по сравнению с базовой моделью. Несмотря на увеличение сложности архитектуры и числа параметров, значения метрик остаются на сопоставимом уровне.

Это свидетельствует о том, что в задаче реконструкции входных данных ключевую роль играет не столько структура модели, сколько информативность входного представления. Разделение потоков обработки оказывается менее эффективным, чем в задачах

классификации или регрессии, поскольку реконструкция не требует явного выделения признаков.

Результаты оценки вычислительной эффективности моделей представлены в табл. 4.

ТАБЛИЦА IV. ВЫЧИСЛИТЕЛЬНАЯ ЭФФЕКТИВНОСТЬ МОДЕЛЕЙ

Конфигурация модели	**Время GPU (мс)**	**Время CPU (мс)**	**Размер модели (МБ)**
Базовая модель	4.8 – 6.2	38 – 52	18 – 25
Модель с прунингом	3.1 – 4.0	25 – 36	10 – 15
Прунинг + квантование + ONNX	1.6 – 2.4	12 – 20	4 – 8

Анализ показывает, что применение методов оптимизации позволяет существенно снизить вычислительные затраты. Прунинг обеспечивает уменьшение времени инференса без значительного ухудшения качества, что делает его наиболее эффективным методом компрессии моделей.

Квантование и оптимизация вычислительного графа обеспечивают дополнительное ускорение, однако могут приводить к ухудшению качества. Наиболее выраженный выигрыш наблюдается при выполнении моделей на центральном процессоре, что делает данные методы особенно актуальными для использования на устройствах с ограниченными ресурсами.

Сопоставление результатов для различных типов моделей позволяет выявить принципиальные различия в их поведении. Сверточные модели демонстрируют устойчивое улучшение качества при использовании дополнительных спектральных признаков, тогда как автоэнкодерные архитектуры характеризуются более сложной и неоднозначной зависимостью.

Данное различие обусловлено постановкой задачи. В задаче классификации модель ориентирована на извлечение наиболее информативных признаков, что делает использование спектральных индексов эффективным. В задаче реконструкции модель стремится восстановить исходное представление, и избыточность признаков может затруднять процесс обучения.

Кроме того, результаты показывают, что увеличение сложности архитектуры не гарантирует улучшения качества. Это подтверждает необходимость тщательного выбора структуры модели с учётом специфики задачи и свойств данных.

IV. ЗАКЛЮЧЕНИЕ

В ходе проведённого исследования выполнен сравнительный анализ сверточных автоэнкодеров и сверточных нейронных сетей, применяемых для обработки многоспектральных данных, с учётом различных методов оптимизации моделей. Рассмотрены подходы, направленные на снижение вычислительной сложности, включая прунинг, квантование и использование оптимизированных форматов представления вычислительных графов.

Полученные результаты показывают, что влияние методов оптимизации существенно зависит от архитектурных особенностей модели и характера решаемой задачи. Для автоэнкодерных архитектур установлено, что уменьшение избыточности параметров способствует повышению обобщающей способности модели, что проявляется в улучшении качества реконструкции. Данный эффект обусловлен снижением степени переобучения за счёт ограничения мощности модели.

Для сверточных нейронных сетей, используемых в задачах классификации и регрессии, выявлена зависимость между степенью оптимизации и точностными характеристиками. Уменьшение сложности модели сопровождается определённым снижением качества предсказания, что отражает наличие компромисса между вычислительной эффективностью и точностью.

Отдельного внимания заслуживает использование стандартизированных форматов представления моделей, обеспечивающих оптимизацию вычислительного графа на уровне среды исполнения. Применение данного подхода позволяет повысить производительность без изменения структуры модели и без дополнительной деградации качества.

В целом результаты исследования подтверждают целесообразность применения методов оптимизации при разработке нейросетевых моделей для обработки многоспектральных данных, особенно в условиях ограниченных вычислительных ресурсов. Выбор конкретного метода оптимизации должен осуществляться с учётом требований к точности и доступных вычислительных возможностей, а также специфики решаемой задачи.

V. ДАЛЬНЕЙШИЕ ПЕРСПЕКТИВЫ ИССЛЕДОВАНИЯ

Проведённое исследование показало эффективность комбинированного подхода, объединяющего спектральные индексы с нейросетевыми архитектурами, а также подтвердило целесообразность применения методов оптимизации для повышения вычислительной эффективности моделей. Вместе с тем в ходе работы был выявлен ряд направлений, требующих дальнейшего развития.

Прежде всего, перспективным является расширение спектрального диапазона за счёт перехода к полноценным гиперспектральным данным с десятками и сотнями каналов, что открывает возможности для более точного анализа типов растительности и почв. Однако высокая размерность требует разработки эффективных методов снижения размерности, в частности вариационных автоэнкодеров и архитектур на основе трансформеров, которые в настоящее время практически не применяются для обработки многоспектральных данных.

Важным направлением является адаптация моделей к изменяющимся условиям съёмки, поскольку спектральный отклик, зарегистрированный спутниковой камерой, отличается от отклика, полученного с беспилотного летательного аппарата, ввиду влияния атмосферных условий, высоты и угла освещения. Разработка методов адаптации доменов и переноса обучения позволит сохранять высокую точность при переходе между разными сенсорами без повторного сбора размеченных данных. Также заслуживает внимания применение трансформерных архитектур

(ViT), способных моделировать как пространственные, так и спектральные взаимосвязи.

В части оптимизации моделей перспективным является использование полного целочисленного квантования или квантования с обучением, что потенциально способно обеспечить дополнительный прирост производительности. Актуальна также разработка специализированных архитектур для встраиваемых платформ (на базе MobileNet или EfficientNet-lite) с учётом специфики многоспектральных данных.

Завершающим этапом должна стать интеграция разработанных моделей в действующие программно-аппаратные комплексы дистанционного зондирования на базе беспилотных летательных аппаратов для оценки эффективности в реальных условиях эксплуатации. Дальнейшее развитие работы будет направлено на повышение точности, устойчивости и вычислительной эффективности методов анализа многоспектральных данных, а также на их практическое внедрение в системы оперативного мониторинга.

Список литературы

[1] Landgrebe D. (2002) Hyperspectral image data analysis. IEEE Signal Process. Mag. 19, P. 17–28.

[2] LeCun Y., Bengio Y., & Hinton G. (2015) Deep learning. Nature. 521, P. 436–444.

[3] Simonyan K., & Zisserman A. (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

[4] LeCun Y., Bottou L., Bengio Y., & Haffner P. (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE. 86, P. 2278–2324.

[5] Chen Y., Lin Z., Zhao X., Wang G., Gu Y. Deep learning-based classification of hyperspectral data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2014, 7, 2094–2107.

[6] Malthus T.J., Mumby P.J. Remote sensing of the coastal zone: An overview and priorities for future research. Int. J. Remote Sens. 2003, 24, 2805–2815.

[7] Hu F., Xia G.S., Hu J., Zhang L. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sens. 2015, 7, 14680.

[8] Lyu H., Lu H., Mou L. Learning a transferable change rule from a recurrent neural network for land cover change detection. Remote Sens. 2016, 8, 506.

[9] Zuo Z., Shuai B., Wang G., Liu X., Wang X., Wang B., Chen Y. Convolutional recurrent neural networks: Learning spatial dependencies for image representation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 18–26.