УДК 004.81

КВАНТОВАНИЕ VISION TRANSFORMER: CPU-ЦЕНТРИЧНЫЙ АНАЛИЗ КОМПРОМИССА МЕЖДУ РАЗМЕРОМ МОДЕЛИ И СКОРОСТЬЮ ИНФЕРЕНСА

А. Р. Нигматуллин[1] [0009-0001-6884-1119], Р. А. Лукманов[2] [0000-0001-9257-7410], А. Таха[3] [0009-0006-6346-4162]

[1–3]*Университет Иннополис, г. Иннополис, Россия*

[1]*Центр искусственного интеллекта Университета Иннополис, г. Иннополис, Россия*

[1]am.nigmatullin@innopolis.university, [2]r.lukmanov@innopolis.university, [3]a.taha@innopolis.university

Аннотация

Использование моделей Vision Transformer (ViT) в реальной медицинской практике, например в больницах или диагностических центрах, часто затруднено, потому что на рабочих компьютерах врачей обычно нет мощных графических процессоров (GPU), а имеющиеся вычислительные ресурсы ограничены. В настоящей работе рассмотрен полный путь практической реализации модели на этапе применения (pipeline инференса), направленный на снижение вычислительных затрат без существенной потери качества.

Предложенный подход объединяет несколько методов оптимизации. Во-первых, использована дистилляция знаний (knowledge distillation) – метод обучения, при котором компактная модель копирует поведение более крупной и точной модели-учителя. Во-вторых, применено экспоненциальное скользящее среднее (Exponential Moving Average, EMA) весов, позволяющее стабилизировать обучение и повысить обобщающую способность модели. В-третьих, исследована посттренировочная квантизация до целочисленного формата INT8 (post-training quantization, PTQ), направленная на уменьшение размера модели и ускорение инференса. Дополнительно рассмотрен упрощенный вариант квантизации совместно с обучением (QAT-lite), при котором эффекты квантизации частично учитываются во время

дообучения модели.

Эксперименты проведены на датасете ISIC, содержащем дерматоскопические изображения кожных новообразований. Оценка качества моделей включает стандартные метрики классификации: точность (accuracy), макроусредненную F1-меру и площадь под ROC-кривой (ROC-AUC). Проанализированы характеристики производительности на центральном процессоре (CPU), включая задержку инференса, пропускную способность, потребление памяти и итоговый размер модели.

Полученные результаты показали, что посттренировочная INT8-квантизация позволяет сохранить качество, близкое к модели в формате FP32, при существенном снижении требований к памяти и вычислительным ресурсам. В то же время использование QAT-lite не демонстрирует устойчивых и воспроизводимых улучшений по сравнению с PTQ.

***Ключевые слова**: Визуальный трансформер (ViT), дистилляция знаний, экспоненциальная скользящая средняя (EMA), посттренировочная квантизация, обучение с учетом квантования.*

ВВЕДЕНИЕ

Визуальные трансформеры (Vision Transformers, ViT) показывают высокие результаты в задачах компьютерного зрения, включая анализ медицинских изображений. Однако их практическое использование в клинических условиях остается ограниченным. Основная причина этого заключается в высокой вычислительной нагрузке, особенно при работе на стандартных центральных процессорах (CPU), которые широко используются в медицинских учреждениях.

В работе рассмотрен практический подход к применению ViT в медицинской визуализации при ограниченных вычислительных ресурсах. Исследован оптимизационный пайплайн, направленный на уменьшение размера модели и ускорение инференса без заметного ухудшения качества. В качестве основного метода использована дистилляция знаний (knowledge distillation), при которой компактная модель обучается на основе выходов более крупной и точной модели. Для стабилизации обучения применено экспоненциальное скользящее среднее (Exponential Moving Average, EMA) весов. Дополнительно использована посттренировочная квантизация

до целочисленного формата INT8, позволяющая снизить требования к памяти и сократить время обработки изображений.

Представлена последовательная оценка этих методов в контексте медицинских задач. Основное внимание уделено балансу между точностью классификации и вычислительной эффективностью моделей при выполнении инференса на CPU.

За последние годы ViT-модели стали широко применяться в анализе медицинских изображений. Обзорные работы показали, что по сравнению со сверточными нейронными сетями такие модели лучше учитывают глобальный контекст изображения, что важно для медицинских данных [1, 2]. Применения ViT в гистопатологии рассмотрена в [3], где также обсуждены существующие ограничения. Более общие обзоры использования трансформеров в медицинской визуализации представлены в [4, 5]. Работы, посвященные сегментации, подчеркивают роль трансформеров в точном анализе структур и границ на медицинских изображениях [6].

Основным ограничением для применения ViT в клинических задачах медицинской визуализации остается их высокая вычислительная стоимость. Одно из направлений исследований связано с разработкой облегченных архитектур [7]. Однако на практике чаще используют методы сжатия уже обученных моделей. К таким методам относится квантизация, которая позволяет уменьшить объем памяти и ускорить вычисления [8–11]. Обобщенный обзор методов сжатия представлен в [12].

Дистилляция знаний является еще одним распространенным подходом к уменьшению размера моделей при сохранении приемлемой точности. В классической работе Хинтона и соавторов [13] было показано, что компактная модель может эффективно обучаться на выходах более сложной модели. В дальнейшем этот подход был расширен и адаптирован для различных сценариев, включая медицинские задачи [14–17].

В целом проводимые исследования в области медицинской визуализации развиваются в двух направлениях: создание более компактных архитектур и применение методов сжатия для адаптации моделей к ограниченным вычислительным условиям. Настоящая работа объединяет эти подходы и оценивает совместное использование дистилляции знаний,

экспоненциального скользящего среднего весов и INT8-квантизации для ViT-моделей, предназначенных для инференса на CPU.

МЕТОДЫ

Для получения компактной, но при этом точной модели мы используем дистилляцию знаний. В рамках этого подхода меньшая нейронная сеть, называемая моделью-студентом, обучается с опорой на выходы более крупной и предварительно обученной модели-учителя. Основная идея заключается в том, что модель-учитель передает модели-студенту не только правильные ответы, но и более богатую информацию о структуре задачи, содержащуюся в распределении выходных вероятностей.

Обучение модели-студента проводится с использованием комбинированной функции потерь. С одной стороны, используется стандартная функция кросс-энтропии, которая измеряет соответствие предсказаний модели-студента истинным меткам классов. С другой стороны, добавляется функция потерь дистилляции, которая поощряет совпадение распределения выходов модели-студента с распределением выходов модели-учителя. Такое сочетание позволяет сохранить высокую точность даже при существенном уменьшении размера модели.

Комбинированная функция потерь имеет следующий вид:

$$\mathrm{LKD} = (1 - \alpha)\mathrm{CE}(z_s, y) + \alpha T^2 \mathrm{KL}(softmax(\tfrac{z_t}{T})||softmax\left(\tfrac{z_s}{T}\right)),$$

где z_t и z_s обозначают логиты модели-учителя и модели-студента соответственно, y – истинные метки классов, CE(·) — функция кросс-энтропии, KL(·‖·) — дивергенция Кульбака–Лейблера, T – температурный параметр, сглаживающий распределение вероятностей, $\alpha \in [0,1]$ — коэффициент, определяющий баланс между вкладом стандартной функции потерь и потерь дистилляции.

Экспоненциальное скользящее среднее весов

Для повышения устойчивости обучения и улучшения обобщающей способности модели применим экспоненциальное скользящее среднее (Exponential Moving Average, EMA) весов. Вместо использования мгновенных значений параметров модели, EMA поддерживает сглаженную версию весов,

которая обновляется постепенно и менее чувствительна к шуму градиентов.

Обновление EMA выполняется по следующему правилу:

$$m_t = \beta m_{t-1} + (1-\beta)\theta_t, \ \ \beta = 0.999,$$

где θ_t – текущие значения весов модели на шаге обучения t, а m_t – соответствующие EMA-веса. Использование EMA позволяет получить более стабильную модель для оценки и инференса, что особенно важно в условиях ограниченных вычислительных ресурсов.

Квантизация модели

Для дальнейшего уменьшения вычислительной нагрузки и объема памяти используют квантизацию параметров модели. В настоящей работе применено аффинное квантизирование, при котором значения с плавающей запятой преобразуются в 8-битные целые числа. Преобразование описывается следующими соотношениями:

$$x_{int} = \mathrm{round}\left(\frac{x}{s}\right) + z, \ \ \hat{x} = s(x_{\mathrm{int}} - z).$$

где s – масштаб, а x_{int} – нулевая точка. Эти параметры подбираются отдельно для каждого слоя модели, что позволяет более точно аппроксимировать исходные значения.

При квантизации модели в целые числа переводят только полносвязные слои, так как они выполняют большинство вычислений. Другие операции, которые отвечают за нормализацию данных внутри сети и преобразование выходов модели в вероятности для классов, оставляют в привычном формате с плавающей запятой (FP32). Это делается потому, что такие операции очень чувствительны к точности чисел, и если их квантизировать полностью, модель может работать нестабильно.

Мы рассматриваем два варианта квантизации. В случае посттренировочной квантизации (Post-Training Quantization, PTQ) модель квантизируется после завершения обучения без изменения весов. В варианте QAT-lite эффекты квантизации частично учитываются во время короткого этапа дообучения за счет использования так называемой «фейковой» квантизации, имитирующей целочисленные вычисления.

Архитектуры и базовые модели

Мы используем подход «учитель – студент», при котором одна модель служит источником знаний, а другая – компактной версией, предназначенной для практического применения. В качестве модели-учителя выбрана архитектура DeiT-Small@224. Это визуальный трансформер, который демонстрирует высокую точность при работе с изображениями стандартного разрешения и широко используется в исследовательских работах как надежная и сбалансированная базовая модель. Его вычислительная сложность делает его удобным эталоном качества, однако в клинических условиях такая модель часто оказывается слишком ресурсоемкой для повседневного использования.

В роли модели-студента была использована DeiT-Tiny@224 – более компактная версия той же архитектуры. По сравнению с моделью-учителя она содержит существенно меньше параметров и требует меньших вычислительных затрат, что делает ее более подходящей для развертывания в средах с ограниченными ресурсами. В частности, такая модель может использоваться для инференса на центральном процессоре без необходимости применения графических ускорителей, что соответствует типичным условиям эксплуатации в медицинских учреждениях.

Для корректной оценки качествоа и практической применимости модели-студента ее характеристики сравниваем не только с моделью-учителем, но и с рядом широко распространенных сверточных нейронных сетей. Эти модели выбраны таким образом, чтобы представить разные поколения и различные подходы в архитектурах для анализа изображений.

В качестве базовой модели была использована ResNet-18, как одна из наиболее часто используемых архитектур в задачах компьютерного зрения. Это одна из наиболее известных и хорошо изученных архитектур, которая часто применяется в медицинской визуализации и служит удобной точкой отсчета при сравнении новых методов. Модель MobileNetV3-Large включена в сравнение как пример архитектуры, специально разработанной для эффективного инференса при ограниченных вычислительных ресурсах. Такие модели широко используют в мобильных и встроенных системах, где важны низкая задержка и малое потребление памяти. Дополнительно

рассмотрена ConvNeXt-Tiny – современная сверточная архитектура, которая заимствует ряд идей из трансформеров и демонстрирует высокое качество при относительно умеренной вычислительной сложности. Эта модель использована в качестве сильного современного ориентира среди CNN.

Таким образом, выбранный набор моделей позволяет оценить положение компактного визуального трансформера относительно как более тяжелых моделей трансформеров, так и различных сверточных архитектур, применяемых на практике.

Оценка вычислительной сложности

При анализе вычислительной сложности моделей мы учитываем не только стандартные теоретические показатели, но и характеристики, важные для реального использования в клинических условиях. В частности, оцениваем общее число параметров модели и теоретическую вычислительную нагрузку, выраженную в количестве операций с плавающей запятой (FLOPs). Эти показатели дают общее представление о сложности архитектуры, однако не всегда отражают реальные затраты при развертывании. Поэтому дополнительно рассматриваем практические метрики, такие как фактический размер контрольной точки модели на диске. Этот показатель напрямую связан с требованиями к хранению данных, скорости загрузки модели и возможностям ее обновления в медицинских информационных системах. Учет как теоретических, так и практических характеристик позволяет более полно оценить пригодность моделей для использования в клинических сценариях с ограниченными вычислительными ресурсами.

ЭКСПЕРИМЕНТЫ

Эксперименты были проведены на наборе данных ISIC – общедоступном медицинском датасете, предоставленном International Skin Imaging Collaboration и содержащем дерматоскопические изображения кожных поражений. Этот набор данных широко используется для оценки алгоритмов автоматической классификации в дерматологии и является стандартным набором в исследованиях по медицинской визуализации. Мы использовали заранее определенные разбиения данных на обучающую, валидационную

и тестовую выборки.

Для обеспечения воспроизводимости экспериментов во всех запусках были применены фиксированные значения случайных инициализаций. На этапе тестирования аугментации изображений не использовались, чтобы полученные результаты отражали поведение моделей в условиях практического применения. Измерения производительности при инференсе на центральном процессоре (CPU) проводились с учетом этапа разогрева, после чего инференс запускался несколько раз подряд. Это позволило получить устойчивые оценки времени обработки одного изображения.

Обучение моделей выполнялось с использованием оптимизатора AdamW. Скорость обучения изменялась по косинусному расписанию, обеспечивающему плавное снижение шага оптимизации. В процессе дистилляции знаний было применено экспоненциальное скользящее среднее (Exponential Moving Average, EMA) весов, что позволило получить более стабильные параметры модели для последующей оценки.

Квантизация моделей была реализована с использованием backend fbgemm, оптимизированного для целочисленных вычислений на CPU. В случае посттренировочной квантизации (Post-Training Quantization, PTQ) применялась динамическая квантизация линейных слоев без дополнительного обучения модели. Для варианта QAT-lite был использован короткий этап дообучения продолжительностью пять эпох, в течение которого эффекты квантизации учитывались во время обучения перед сохранением итоговых весов модели.

Качество классификации было оценено с использованием стандартных метрик: точности (accuracy), макроусредненной F1-меры и площади под ROC-кривой (ROC-AUC), которая отражает способность модели различать классы при разных порогах принятия решения. Для оценки практической применимости моделей в клинических условиях дополнительно были проанализированы вычислительные характеристики при работе на CPU, включая задержку инференса, пропускную способность, пиковое использование оперативной памяти (RAM) и итоговый размер модели на диске. Совокупность этих показателей позволяет оценить как качество предсказаний, так и вычислительную пригодность моделей для использования в реальных медицинских сценариях.

РЕЗУЛЬТАТЫ

Далее представлены результаты сравнения моделей по качеству классификации, требованиям к памяти и производительности при инференсе на центральном процессоре (Intel i7-12700F). Основное внимание уделено влиянию посттренировочной INT8-квантизации и ее сочетания с дистилляцией знаний и экспоненциальным скользящим средним весов.

Качество классификации и использование памяти

Эксперименты на наборе данных ISIC показали, что посттренировочная INT8-квантизация практически не ухудшает качество дистилляционной модели DeiT-Tiny по сравнению с базовой версией в формате с плавающей запятой (FP32). На валидационной выборке точность изменилась всего на − 0.13 п. п., а макроусредненная F1-мера даже незначительно выросла (+0.27 п.п.). Аналогичная картина наблюдалась и на тестовом разбиении, где изменения составили −0.10 п.п. по точности и +0.08 п.п. по макро-F1. При этом выигрыш в компактности модели оказался существенным. Размер модели на диске уменьшился с 21.13 до 5.97 МБ, то есть примерно в 3.5 раза (−71.7%). Пиковое использование оперативной памяти во время инференса также снизилось примерно на 247 МБ, что соответствует уменьшению на 14%. Эти результаты особенно важны для клинических сценариев, где ограничения по памяти и хранению данных часто являются критичными.

Задержка и пропускная способность на CPU

Производительность моделей на CPU зависит от размера обрабатываемого пакета изображений (batch size) и используемого варианта модели. При обработке одного изображения за раз (batch = 1) квантизированная модель Student INT8 (PTQ) оказалась немного медленнее версии FP32: медианная задержка составила 16.49 мс против 14.53 мс, а пропускная способность снизилась примерно на 11%. Это связано с дополнительными накладными расходами на операции квантования и деквантования. Однако при использовании экспоненциального скользящего среднего весов (KD+EMA+PTQ) эта разница практически исчезла. В данном случае задержка инференса почти совпала с лучшим вариантом FP32

и оказалась заметно ниже, чем у модели KD+EMA в формате FP32 (14.46 мс против 16.77 мс).

При увеличении размера пакета до batch = 8 негативный эффект квантизации исчезает. Простая PTQ-квантизация показала задержку на уровне FP32 (58.45 мс против 58.76 мс), а сочетание KD+EMA+PTQ продемонстрировало преимущество по скорости по сравнению с KD+EMA FP32 (53.38 мс против 59.22 мс), а также более высокую пропускную способность (+8.2%). Это указывает на то, что квантизация особенно эффективна при вычислительно нагруженных сценариях.

Сравнение с базовыми CNN-архитектурами

Сравнение с распространенными сверточными архитектурами показало, что дистилляционный DeiT-Tiny остается конкурентоспособным при инференсе на CPU. При batch = 1 модель DeiT-Tiny FP32 (KD) работает быстрее, чем ResNet-18 FP32 (14.53 мс против 15.29 мс), и значительно быстрее, чем ConvNeXt-Tiny FP32 (38.28 мс). При batch = 8 архитектура MobileNetV3-Large FP32 сохраняет лидерство по пропускной способности, что ожидаемо с учетом ее ориентации на высокоэффективный инференс.

Вариант квантизации с учетом обучения (QAT-lite) не продемонстрировал устойчивых преимуществ. Он уступает PTQ по качеству классификации и не обеспечивает заметного выигрыша по задержке ни при одном из рассмотренных размеров пакета.

Практические выводы

С точки зрения практического развертывания полученные результаты показали, что посттренировочная INT8-квантизация является простым и надежным способом уменьшить размер ViT-моделей примерно в 3.5 раза при сохранении качества практически на уровне FP32. Такой подход особенно полезен в сценариях, где ограничения по памяти, хранению данных или распространению моделей играют ключевую роль.

Дополнительно эти результаты позволяют четко определить условия, при которых квантизация дает наибольший эффект. Ускорение наблюдается в вычислительно нагруженных режимах, при использовании более крупных пакетов изображений и в вариантах с EMA-стабилизацией весов. В то же время

задержка обработки одного изображения остается ограниченной накладными расходами на операции квантования и деквантования вокруг чувствительных слоев, таких как Layer Normalization и Softmax. Подробные численные результаты сравнения представлены в табл. 1, 2.

Динамика обучения модели-студента проанализирована в зависимости от номера эпохи (*epoch*). Под одной эпохой понимается один полный проход всей обучающей выборки через модель с последующим обновлением параметров. На рис. 1 показаны кривые для обучающей выборки при использовании дистилляции знаний и экспоненциального скользящего среднего весов. Видно, что применение EMA способствует более стабильному и плавному снижению значения функции потерь, а также уменьшает флуктуации точности в процессе обучения.

Аналогичный анализ на валидационной выборке представлен на рис. 2. В этом случае также наблюдается отсутствие резких скачков метрик, что указывает на лучшую обобщающую способность модели и снижение риска переобучения.

Отдельно рассмотрен короткий этап дообучения с учетом квантизации (QAT-lite). Соответствующие кривые представлены на рис. 3. Можно отметить, что несмотря на небольшое улучшение сходимости на первых итерациях, дальнейшее обучение не приводит к существенному росту качества, что согласуется с результатами количественного сравнения и подтверждает ограниченную эффективность QAT-lite по сравнению с посттренировочной квантизацией.

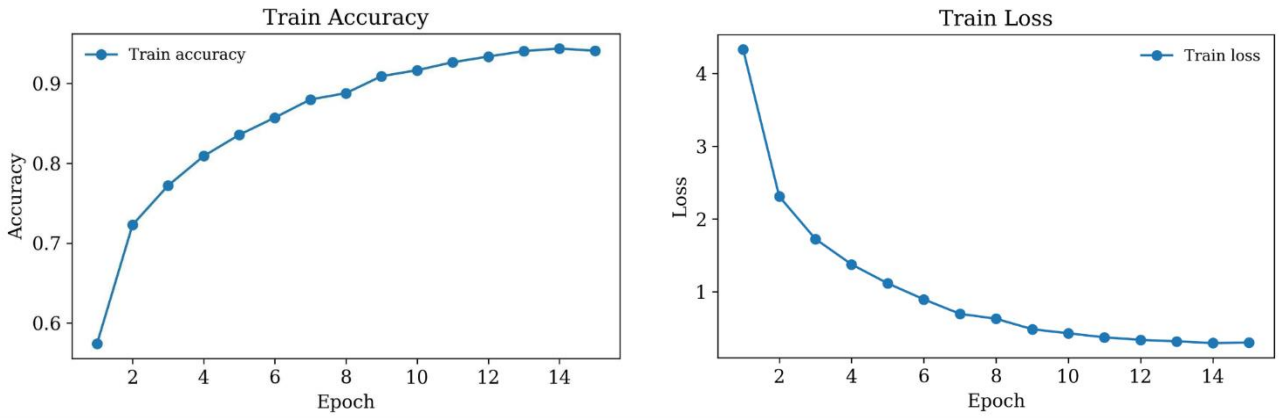

Рис. 1. Кривые изменения функции потерь (loss) и точности (accuracy) во время обучения модели-студента DeiT-Tiny с применением дистилляции знаний (KD) и экспоненциального скользящего среднего весов (EMA).

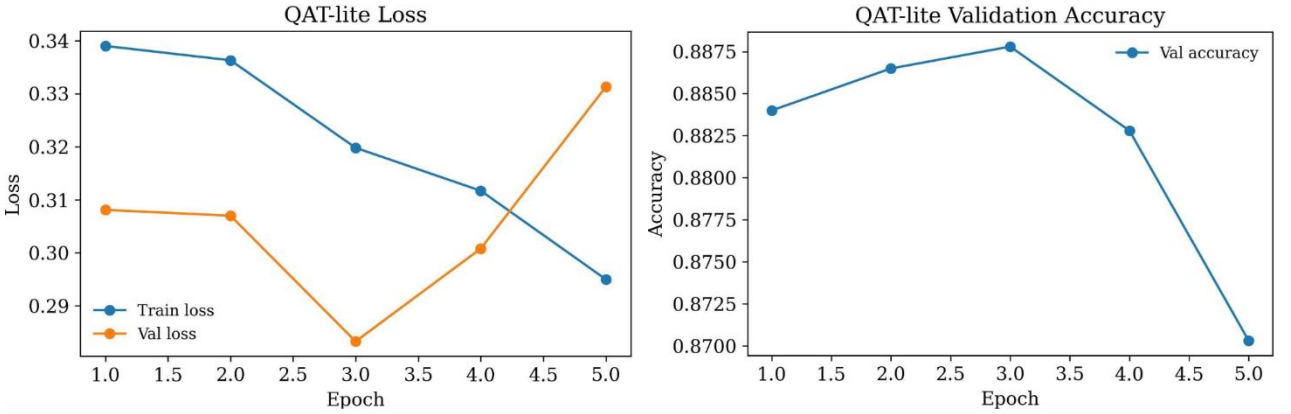

Рис. 2. Кривые изменения функции потерь и точности на валидационной выборке в процессе обучения модели-студента DeiT-Tiny.

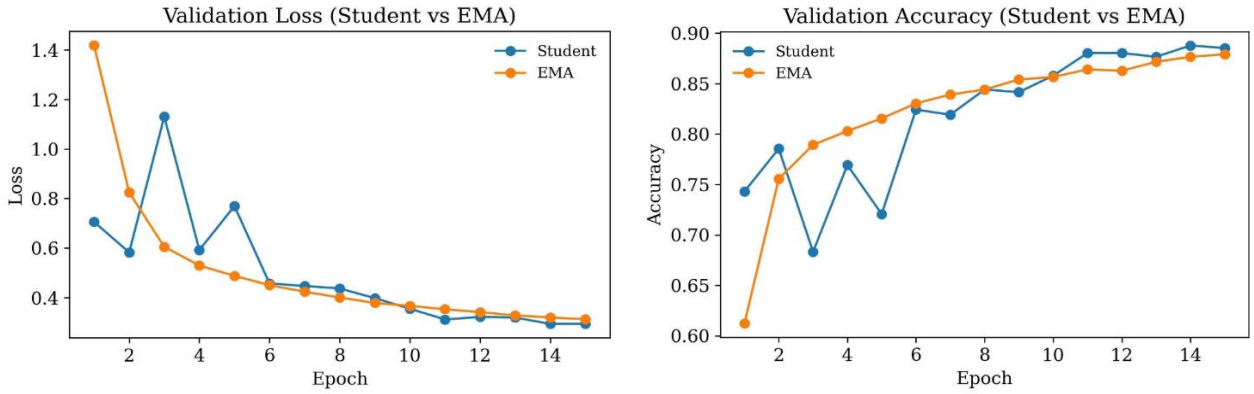

Рис. 3. Кривые изменения метрик модели во время короткого этапа дообучения с учетом квантизации (QAT-lite). Этот этап включает как дистилляцию знаний (KD), так и имитацию целочисленных вычислений во время обучения, чтобы модель могла корректно работать в формате INT8.

Для оценки влияния различных подходов к обучению и упрощению числовых представлений была выполнена экспериментальная проверка качества моделей на валидационной и тестовой выборках.

Результаты получены на наборе данных ISIC, который состоит из медицинских изображений кожи, используемых для задачи классификации кожных поражений. Для экспериментов были применены заранее зафиксированные разбиения данных на обучающую, валидационную и тестовую части. При оценке качества модели изображения использовались

без дополнительных преобразований, а все эксперименты выполнялись с фиксированными начальными условиями, что обеспечивает воспроизводимость результатов.

Табл. 1. Метрики валидации и теста.

Модель	Val Acc	Val Macro-F1	Val ROC-AUC	Test Acc	Test Macro-F1	Test ROC-AUC
Student FP32 (KD)	0.8878	0.8874	0.9903	0.8333	0.7857	0.9658
Student FP32 (KD+EMA)	0.8791	0.8787	0.9889	0.8367	0.7903	0.9657
Student INT8 (PTQ)	0.8865	0.8901	0.9901	0.8323	0.7865	0.9658
Student INT8 (KD+EMA+PTQ)	0.8741	0.8700	0.9891	0.8342	0.7908	0.9652
Student QAT-lite FP32	0.8678	0.8789	0.9908	0.8103	0.7834	0.9634
Student QAT-lite INT8	0.8666	0.8769	0.9907	0.8113	0.7897	0.9631

В табл. 1 представлены значения основных метрик качества для компактной модели DeiT-Tiny@224, обученной с использованием более крупной и точной модели DeiT-Small@224 в качестве источника знаний. Такой подход позволил уменьшить размер модели при сохранении высокой точности. В ряде экспериментов применялось экспоненциальное скользящее среднее весов (EMA) с коэффициентом $\beta = 0.999$, которое сглаживает изменения параметров модели и делает итоговые предсказания более стабильными. Обозначение INT8 (PTQ) соответствует упрощению числовых представлений параметров модели после завершения обучения, что уменьшает ее размер и требования к памяти. При этом наиболее чувствительные вычислительные операции сохраняются в исходном формате, чтобы избежать нестабильности вычислений. Вариант QAT-lite включает короткий этап дополнительного дообучения модели, во время которого она подстраивается под последующее упрощение числовых представлений.

Качество моделей было оценено на валидационной (Val) и тестовой (Test) выборках по следующим показателям:

1) Accuracy – доля правильных предсказаний;

2) Macro-F1 – усредненная мера качества по всем классам, одинаково учитывающая каждый из них;

3) ROC-AUC – показатель способности модели различать классы на основе предсказанных вероятностей.

Во всех случаях более высокие значения метрик соответствуют лучшему качеству модели.

Сравнение методов уменьшения модели

В табл. 2 представлены результаты сравнения степени уменьшения размера различных вариантов модели DeiT-Tiny, обученных и оптимизированных различными способами, на наборе медицинских изображений ISIC. Во всех случаях была использована одна и та же архитектура модели, поэтому варианты не отличаются по своей вычислительной сложности: число параметров составляет 5.53 млн, а объем вычислений – 2.15 млрд операций для изображений с разрешением 224 × 224.

Табл. 2. Сводные по степени сжатия различных вариантов модели DeiT-Tiny

Модель	Точность	Размер (МБ)	Compression factor
DeiT-T FP32 (KD)	FP32	21.13	1.00
DeiT-T FP32 (KD+EMA)	FP32	21.13	1.00
DeiT-T INT8 (PTQ)	INT8	5.97	3.54
DeiT-T INT8 (KD+EMA+PTQ)	INT8	5.97	3.54
DeiT-T QAT-lite FP32	FP32	21.13	1.00
DeiT-T QAT-lite INT8	INT8	5.97	3.54

Отличия данных для различных моделей связаны исключительно со способом представления параметров модели. Варианты с пометкой INT8 используют упрощенный числовой формат для части операций, что позволяет существенно сократить занимаемое пространство. При этом упрощение применяется только к основным линейным слоям модели, тогда как наиболее чувствительные операции, отвечающие за нормализацию и вычисление вероятностей, сохраняются в исходном формате для обеспечения стабильной работы. Указанные в таблице размеры моделей соответствуют фактическому объему файлов с сохраненными весами на диске, что напрямую отражает требования к хранению и передаче модели в реальных сценариях развертывания.

С учетом указанных различий в представлении параметров ниже приведены результаты измерения производительности моделей на CPU.

CPU benchmarks

После анализа степени сжатия и размера моделей рассмотрим их практическую производительность при инференсе на центральном процессоре (CPU).

Табл. 3. Производительность моделей на CPU при обработке одного изображения за раз (batch = 1). Более низкая задержка и меньший расход памяти, а также более высокая пропускная способность означают лучшую эффективность.

Модель	p50 (мс)	p90 (мс)	Throughput (кол-во/с)	Peak RAM (МБ)	Size (MB)
DeiT-T FP32 (KD)	14.53	15.91	67.7	1765.1	21.13
DeiT-T FP32 (KD+EMA)	16.77	17.87	59.2	1779.0	21.13
DeiT-T INT8 (PTQ)	16.49	17.17	60.3	1518.3	5.97
DeiT-T INT8 (KD+EMA+PTQ)	14.46	16.84	66.9	1532.2	5.97
DeiT-T QAT-lite FP32	17.26	18.34	58.0	1539.8	21.13
DeiT-T QAT-lite INT8	15.77	17.55	62.3	1533.2	5.97
ResNet-18 FP32	15.29	17.69	61.9	1539.6	42.72
MobileNetV3-L FP32	16.58	17.19	60.4	1616.1	16.25
ConvNeXt-T FP32	38.28	39.21	26.4	1642.2	106.21
ResNet-18 INT8 (PTQ)	17.54	19.04	56.3	1675.7	42.71
MobileNetV3-L INT8 (PTQ)	17.56	18.06	56.7	1675.7	16.25
ConvNeXt-T INT8 (PTQ)	36.16	36.82	27.8	1675.3	32.18

Измерения производительности выполнены по следующему протоколу: 50 итераций использованы для разогрева системы, после этого выполнены 100 синхронизированных измерительных итераций; весь процесс повторялся пять раз для повышения надежности результатов. Показатели, приведенные в табл. 3, интерпретируются следующим образом:

1. p50/p90 – медианное и 90-й перцентиль времени обработки одного изображения (в миллисекундах);
2. Throughput – количество изображений, обрабатываемых моделью в секунду в ходе измерений;
3. Peak RAM – максимальный объем оперативной памяти, используемый во время инференса (в мегабайтах);

4. Size – фактический размер файла с сохраненными весами модели на диске (в мегабайтах).

Все варианты модели DeiT-Tiny используют одинаковую архитектуру, следовательно, имеют одинаковую теоретическую вычислительную сложность: 5.53 млн параметров и 2.15 млрд операций для изображений с разрешением 224 × 224. В вариантах с INT8-квантизацией упрощение числовых представлений было применено только к линейным слоям модели с использованием библиотеки fbgemm, тогда как операции нормализации и вычисления вероятностей сохранялись в исходном формате для обеспечения стабильности вычислений. В табл. 4 представлены экспериментальные показатели производительности различных вариантов модели на CPU.

Табл. 4. Производительность моделей на CPU при batch = 8 (50 итераций разогрева, 100 измерительных итераций, 5 повторов)

Модель	p50 (мс)	p90 (мс)	Throughput (кол-во/с)	Peak RAM (МБ)	Size (MB)
DeiT-T FP32 (KD)	58.76	60.45	137.1	1779.0	21.13
DeiT-T FP32 (KD+EMA)	59.22	61.11	136.1	1779.0	21.13
DeiT-T INT8 (PTQ)	58.45	61.16	136.2	1528.5	5.97
DeiT-T INT8 (KD+EMA+PTQ)	53.38	59.25	147.3	1538.6	5.97
DeiT-T QAT-lite FP32	59.45	60.75	134.2	1539.8	21.13
DeiT-T QAT-lite INT8	57.79	59.75	138.2	1538.2	5.97
ResNet-18 FP32	78.82	81.17	101.5	1613.9	42.72
MobileNetV3-L FP32	48.04	48.58	166.4	1641.3	16.25
ConvNeXt-T FP32	189.68	193.84	42.0	1675.6	106.21
ResNet-18 INT8 (PTQ)	78.56	80.60	101.5	1675.7	42.71
MobileNetV3-L INT8 (PTQ)	48.25	48.84	165.4	1675.7	16.25
ConvNeXt-T INT8 (PTQ)	162.62	165.86	49.2	1714.2	32.18

Измерения производительности выполнены согласно протоколу, описанному для табл. 3.

В табл. 5 приведены результаты абляционных экспериментов для различных вариантов компактной модели DeiT-Tiny на валидационной выборке при инференсе на CPU с размером пакета batch = 1.

Все варианты модели используют ту же самую архитектуру (5.53 млн параметров, 2.15 млрд операций при разрешении 224 × 224). Варианты с INT8-квантизацией (PTQ или QAT-lite INT8) упрощают представление чисел только для линейных слоев с использованием библиотеки fbgemm, при этом операции LayerNorm и Softmax были сохранены в исходном формате для числовой стабильности.

Табл. 5. Абляционные эксперименты (валидационная выборка; CPU, batch=1)

Variant	Acc	Macro-F1	ROC-AUC	p50 (мс)	Thr (кол-во/с)	Size (МБ)
KD (Student FP32)	0.8878	0.8874	0.9903	14.53	67.7	21.13
EMA (KD+EMA FP32)	0.8791	0.8787	0.9889	16.77	59.2	21.13
PTQ (KD INT8)	0.8865	0.8901	0.9901	16.49	60.3	5.97
KD+EMA+PTQ	0.8741	0.8700	0.9891	14.46	66.9	5.97
QAT-lite FP32	0.8678	0.8789	0.9908	17.26	58.0	21.13
QAT-lite INT8	0.8666	0.8769	0.9907	15.77	62.3	5.97

ОБСУЖДЕНИЕ

Проведенные эксперименты показали, что сочетание дистилляции знаний и посттренировочной квантизации является эффективным способом использования моделей Vision Transformer (ViT) на обычных CPU, особенно при ограничениях по памяти. Применение INT8-квантизации к линейным слоям модели DeiT-Tiny позволило уменьшить размер модели почти в 3.5 раза – с 21.13 до 5.97 МБ, при этом точность осталась практически на прежнем

уровне. Это говорит о том, что значительное сжатие возможно без сложного дополнительного обучения.

Установлено, что меньшая модель на CPU не всегда работает быстрее. При обработке одного изображения за раз (batch = 1) модели с INT8-квантизацией иногда даже немного медленнее, чем исходные FP32-модели. Причиной служат накладные расходы на преобразование чисел между INT8 и FP32 в слоях LayerNorm и Softmax, которые компенсируют преимущества целочисленных вычислений. Существенное улучшение пропускной способности наблюдается только при пакетной обработке нескольких изображений (batch = 8), что подтверждает, что INT8 выгоден в вычислительно насыщенных сценариях, но не для одиночных картинок.

Таким образом, посттренировочная квантизация PTQ оказалась простым и безопасным методом уменьшения модели без потери точности, особенно когда важна экономия памяти. Однако если приоритет – минимальная задержка при обработке одного изображения, преимущества квантизации ограничены. Кроме того, необходимо учитывать особенности конкретного процессора, так как на других CPU результаты могут отличаться.

ЗАКЛЮЧЕНИЕ

Мы показали, что при использовании моделей Vision Transformer в медицинских приложениях на обычных CPU квантизация не всегда дает ускорение. Для успешного развертывания таких моделей важно учитывать не только методы сжатия, но и особенности инференса на конкретном процессоре. Наши результаты подчеркивают, что подходы, которые уменьшают размер модели, не всегда автоматически сокращают задержку при обработке отдельных изображений.

В будущем перспективными представляются методы, нацеленные именно на снижение задержки, такие как структурная обрезка отдельных блоков модели, схемы квантизации, учитывающие особенности аппаратуры, выборочная квантизация для снижения накладных расходов, а также изучение влияния сжатия на точность калибровки и неопределенность предсказаний. Все это поможет повысить надежность и доверие к медицинским системам на основе искусственного интеллекта.

Благодарности

Работа поддержана Академией наук Республики Татарстан, договор №254/2024-PD.

СПИСОК ЛИТЕРАТУРЫ

1. *Shamshad F., Khan S., Zamir S.W., et al.* Transformers in Medical Imaging: A Survey // arXiv. 2022.

2. *He K., Gan C., et al.* Transformers in Medical Image Analysis: A Review // arXiv. 2022.

3. *Atabansi C.C., Nie J., et al.* A Survey of Transformer Applications for Histopathological Image Analysis: New Developments and Future Directions // Biomedical Engineering Online. 2023. Vol. 22, No. 1. https://doi.org/10.1186/s12938-023-01069-5

4. *Azad R., Kazerouni A., Heidari M., et al.* Advances in Medical Image Analysis with Vision Transformers: A Comprehensive Review // arXiv. 2023.

5. *Shamshad F., Khan S., Zamir S.W., et al.* Transformers in Medical Imaging: A Survey // Medical Image Analysis. 2024. Vol. 88. https://doi.org/10.1016/j.media.2023.102843

6. *Liu Y., et al.* A Recent Survey of Vision Transformers for Medical Image Segmentation // *arXiv.* 2023.

7. *Wu F., et al.* Lite Transformer with Long-Short Range Attention // Proceedings of the International Conference on Learning Representations (ICLR). 2020.

8. *Jacob B., et al.* Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018. P. 2704–2713. https://doi.org/10.1109/CVPR.2018.00286

9. *Nagel M., et al.* A White Paper on Neural Network Quantization // arXiv. 2021.

10. *Han S., Mao H., Dally W.J.* Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding // arXiv. 2016.

11. *Yao Z., et al.* ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers // Advances in Neural Information

Processing Systems (NeurIPS). 2022. Vol. 35.

12. *Wikipedia contributors.* Model Compression // Wikipedia. 2025.

13. *Hinton G., Vinyals O., Dean J.* Distilling the Knowledge in a Neural Network // arXiv. 2015.

14. *Gou J., et al.* Knowledge Distillation: A Survey // International Journal of Computer Vision. 2021. Vol. 129, No. 6. P. 1789–1819. https://doi.org/10.1007/s11263-021-01453-z

15. *Umirzakova S., et al.* Simplified Knowledge Distillation for Deep Neural Networks: Bridging the Performance Gap with a Novel Teacher–Student Architecture // Electronics. 2024. Vol. 13, No. 3. https://doi.org/10.3390/electronics13030512

16. *Liang P., et al.* Data-Free Knowledge Distillation with Feature Synthesis and Spatial Consistency for Image Analysis // Scientific Reports. 2024. Vol. 14, No. 1. https://doi.org/10.1038/s41598-024-53241-3

VIT QUANTIZATION: CPU-CENTRIC ANALYSIS OF THE TRADE-OFF BETWEEN SIZE AND SPEED

A. R. Nigmatullin[1] [0009-0001-6884-1119], R. A. Lukmanov[2] [0000-0001-9257-7410],

A. Taha[3] [0009-0006-6346-4162]

[1–3]*Innopolis University, Innopolis, Russia*

[1]*The Center of Artificial Intelligence of Innopolis University, Innopolis, Russia*

[1]am.nigmatullin@innopolis.university, [2]r.lukmanov@innopolis.university,
[3]a.taha@innopolis.university

Abstract

Using Vision Transformer (ViT) models in real medical practice – for example, in hospitals or diagnostic centers – is often difficult because doctors' work computers usually do not have powerful graphics processors (GPUs), and computing resources are limited. This work investigates a complete practical pipeline for model inference, aimed at reducing computational costs without significant loss of predictive performance.

The proposed approach combines several optimization techniques. First, knowledge distillation (KD) is used, where a compact student model learns to mimic the behavior of a larger, more accurate teacher model. Second, Exponential Moving Average (EMA) of the model weights is applied to stabilize training and improve generalization. Third, post-training INT8 quantization (PTQ) is explored to reduce model size and accelerate inference. Additionally, a simplified quantization-aware training variant (QAT-lite) is considered, where the effects of quantization are partially incorporated during fine-tuning.

Experiments are conducted on the ISIC dataset, which contains dermoscopic images of skin lesions. Model performance is evaluated using standard classification metrics, including accuracy, macro-averaged F1 score, and area under the ROC curve (ROC-AUC). CPU performance is also analyzed, including inference latency, throughput, memory consumption, and the final model size.

The results show that post-training INT8 quantization preserves performance close to the FP32 baseline while substantially reducing memory and computational requirements. In contrast, QAT-lite does not consistently provide reproducible improvements over PTQ.

***Keywords**: Vision Transformer, knowledge distillation, EMA, post-training quantization, quantization-aware training.*

REFERENCES

1. *Shamshad F., Khan S., Zamir S.W., et al.* Transformers in Medical Imaging: A Survey // arXiv. 2022.

2. *He K., Gan C., et al.* Transformers in Medical Image Analysis: A Review // arXiv. 2022.

3. *Atabansi C.C., Nie J., et al.* A Survey of Transformer Applications for Histopathological Image Analysis: New Developments and Future Directions // Biomedical Engineering Online. 2023. Vol. 22, No. 1.
https://doi.org/10.1186/s12938-023-01069-5

4. *Azad R., Kazerouni A., Heidari M., et al.* Advances in Medical Image Analysis with Vision Transformers: A Comprehensive Review // arXiv. 2023.

5. *Shamshad F., Khan S., Zamir S.W., et al.* Transformers in Medical Imaging: A Survey // Medical Image Analysis. 2024. Vol. 88.

https://doi.org/10.1016/j.media.2023.102843

6. *Liu Y., et al.* A Recent Survey of Vision Transformers for Medical Image Segmentation // arXiv. 2023.

7. *Wu F., et al.* Lite Transformer with Long-Short Range Attention // Proceedings of the International Conference on Learning Representations (ICLR). 2020.

8. *Jacob B., et al.* Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018. P. 2704–2713. https://doi.org/10.1109/CVPR.2018.00286

9. *Nagel M., et al.* A White Paper on Neural Network Quantization // arXiv. 2021.

10. *Han S., Mao H., Dally W.J.* Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding // arXiv. 2016.

11. *Yao Z., et al.* ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers // Advances in Neural Information Processing Systems (NeurIPS). 2022. Vol. 35.

12. *Wikipedia contributors.* Model Compression // *Wikipedia*. 2025.

13. *Hinton G., Vinyals O., Dean J.* Distilling the Knowledge in a Neural Network // arXiv. 2015.

14. *Gou J., et al.* Knowledge Distillation: A Survey // International Journal of Computer Vision. 2021. Vol. 129, No. 6. P. 1789–1819. https://doi.org/10.1007/s11263-021-01453-z

15. *Umirzakova S., et al.* Simplified Knowledge Distillation for Deep Neural Networks: Bridging the Performance Gap with a Novel Teacher–Student Architecture // Electronics. 2024. Vol. 13, No. 3. https://doi.org/10.3390/electronics13030512

16. *Liang P., et al.* Data-Free Knowledge Distillation with Feature Synthesis and Spatial Consistency for Image Analysis // Scientific Reports. 2024. Vol. 14, No. 1. https://doi.org/10.1038/s41598-024-53241-3

СВЕДЕНИЯ ОБ АВТОРАХ

НИГМАТУЛЛИН Амир Рамисович – студент 4 курса Университета Иннополис по направлению «Искусственный интеллект», специализируется на оптимизации моделей трансформеров. Научные интересы включают эффективные архитектуры глубокого обучения, компьютерное зрение, объяснимый ИИ и обучение с подкреплением. Выпускная квалификационная работа посвящена анализу и тестированию методов оптимизации трансформеров с целью повышения эффективности и снижения вычислительных затрат. Победитель хакатона по генерации интерьеров компании Leroy Merlin с проектом в области ИИ для дизайна и визуализации пространств.

Amir Ramisovich NIGMATULLIN – 4th year student at Innopolis University with a degree in Artificial Intelligence, he specializes in optimizing transformer models. His research interests include effective deep learning architectures, computer vision, explicable AI, and reinforcement learning.The final thesis is devoted to the analysis and testing of transformer optimization methods in order to increase efficiency and reduce computational costs. The winner of the hackathon on interior generation by Leroy Merlin with a project in the field of AI for the design and visualization of spaces.

email: am.nigmatullin@innopolis.university
ORCID: 0009-0001-6884-1119

ЛУКМАНОВ Рустам Абубакирович (PhD, Бернский университет, 2021) – научный сотрудник, доцент, специализирующийся на машинном обучении, биоинформатике, анализе данных и объяснимом ИИ. Лауреат награды «Молодые лидеры БРИКС и ШОС» (2023). Преподает курсы по объяснимому ИИ и представлению знаний в Университете Иннополис.

Rustam Abubakirovich LUKMANOV (PhD, University of Bern, 2021) - is a Researcher and Associate Professor specializing in machine learning, bioinformatics, data analysis and explicable AI. Winner of the BRICS and SCO Young Leaders Award (2023). Teaches courses on explicable AI and knowledge representation at Innopolis University.

email: r.lukmanov@innopolis.university
ORCID: 0000-0001-9257-7410

ТАХА Ахмад – аспирант и научный сотрудник Центра искусственного интеллекта в Университете Иннополис. Специализируется на медицинском ИИ, самообучении (SSL) и компьютерном зрении. Его научные интересы также включают обработку естественного языка (NLP) и трансформеры. Является преподавателем на факультете ИИ.

Ahmad TAHA – is a PhD student and Researcher at the Center of Artificial Intelligence, Innopolis University. He specializes in Medical AI, Self-Supervised Learning (SSL), and Computer Vision. His research interests also include Natural Language Processing (NLP) and Transformers. He is an instructor in the AI department.

email: a.taha@innopolis.university

ORCID: 0009-0006-6346-4162

Материал поступил в редакцию 10 ноября 2025 года