УДК 004.8

ГДЕ НАХОДЯТСЯ ЛУЧШИЕ ПРИЗНАКИ? ПОСЛОЙНЫЙ АНАЛИЗ СЛОЕВ ТРАНСФОРМЕРА ДЛЯ ЭФФЕКТИВНОЙ КЛАССИФИКАЦИИ ЭНДОСКОПИЧЕСКИХ ИЗОБРАЖЕНИЙ

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

[1, 2]Университет Иннополис, г. Иннополис, Россия

[1]Центр искусственного интеллекта университета Иннополис, г. Иннополис, Россия

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

Аннотация

В поисках путей развития медицинского искусственного интеллекта показано, что предварительно обученный Vision Transformer с линейным классификатором может достигать высокой и конкурентоспособной производительности в классификации эндоскопических изображений. Представлен систематический послойный анализ, который выявляет источник наиболее важных признаков, оспаривая общепринятую эвристику использования только последнего слоя. Установлен отчетливый феномен «пика перед концом», когда поздне-промежуточный слой предлагает более обобщаемое представление для последующей медицинской задачи. На стандартных наборах данных Kvasir и HyperKvasir предложенный подход с малым количеством параметров не только получить достаточно высокую точность, но и значительно сокращает вычислительные затраты. Полученные работы могут быть рекомендованы в качестве практического руководства по эффективному использованию признаков общих базовых моделей в клинических условиях.

Ключевые слова*: классификация эндоскопических изображений, замороженный кодировщик, извлечение признаков, послойный анализ, визуальный трансформер (ViT), перенос обучения, самоконтролируемое обучение (SSL), медицинский искусственный интеллект.*

ВВЕДЕНИЕ

Эндоскопия желудочно-кишечного тракта (ЖКТ) является краеугольным камнем в диагностике и лечении широкого спектра заболеваний, от воспалительных заболеваний кишечника (ВЗК) до предотвращения колоректального рака путем удаления предраковых полипов [1–3]. Однако эффективность эндоскопии ограничена человеческой интерпретацией, при этом частота пропущенных аденом во время колоноскопии достигает 26% [4, 5]. Для снижения таких диагностических ошибок в качестве перспективного решения появились системы автоматизированной диагностики (САД), основанные на искусственном интеллекте (ИИ) [6].

Современное состояние в эндоскопической САД представлено моделями глубокого обучения (ГО), в частности сверточными нейронными сетями (СНС) [7], такими как ResNet [8], а в последнее время и Vision Transformers (ViT) [9]. Типичная методология применения этих моделей — это полное дообучение (full fine-tuning), при котором модель, предварительно обученная на крупномасштабном наборе данных (например, ImageNet [10]), адаптируется к эндоскопической задаче путем переобучения всех ее параметров или обучения с нуля на специфических эндоскопических наборах данных. Хотя это и дает хорошие результаты, практическое внедрение является серьезным препятствием. Полное дообучение и обучение с нуля требуют значительных ресурсов ГП и длительного времени обучения, что создает барьер для исследовательских и клинических учреждений [11]. Кроме того, эти модели требуют большого количества данных, а стоимость приобретения больших наборов медицинских данных, аннотированных экспертами, является еще одной проблемой в области анализа медицинских изображений [12].

Для преодоления этих проблем в данной работе исследована более эффективная парадигма: использование предварительно обученной модели в качестве фиксированного экстрактора признаков. В этом подходе глубокий кодировщик остается замороженным, а обучается только простой, легковесный неглубокий декодер на высокоуровневых признаках, извлеченных из кодировщика. Этот метод значительно сокращает количество обучаемых параметров и уменьшает время обучения с часов или дней до минут, а также решает проблему нехватки

данных. Успех этого подхода основан на предположении, что крупномасштабные, предварительно обученные кодировщики с богатыми, обобщаемыми признаками достаточно натренированы для решения последующих медицинских задач без дальнейшей модификации.

Эта парадигма ставит два фундаментальных вопроса:

1) Может ли вычислительно простая модель, состоящая из фиксированного кодировщика и неглубокого декодера, достичь или даже превзойти производительность сложных, полностью дообученных систем в классификации эндоскопических изображений?

2) Если да, то где в этом кодировщике находятся лучшие признаки для этой задачи?

Хотя выбор признаков является известной техникой, обычно оно выполняется из последнего слоя, и систематический анализ качества признаков по всей глубине сети отсутствовал. Наша основная гипотеза заключается в том, что оптимальное представление признаков находится не в последнем слое, а в промежуточном слое i^*. Мы можем формально определить это как задачу оптимизации, где мы стремимся найти индекс слоя i^*, который минимизирует потери на валидации $\mathcal{L}_{\text{val}}$ для декодера $g_{\theta_i^*}$, обученного на признаках из этой слоя:

$$i^* = \arg\min_{i \in \{1,\dots,N\}} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{val}}}\left[\mathcal{L}_{\text{CE}}\left(g_{\theta_i^*}\left(\mathbf{h}_i(x)\right), y\right)\right],$$

где $\mathbf{h}_i(x)$ – вектор признаков из слоя i, θ_i^* – оптимальные веса декодера для этого слоя.

Целью настоящей работы являются:

- **достижение высокой производительности с эффективной моделью:** мы демонстрируем, что простой классификатор, обученный на признаках из оптимального слоя замороженного кодировщика, достигает отличных результатов на наборах данных HyperKvasir и других бенчмарках;
- **новый послойный анализ признаков:** мы представим, насколько нам известно, первый систематический послойный анализ качества признаков из замороженного кодировщика для эндоскопической классификации на стандартных отраслевых бенчмарках;

- **количественная оценка эффективности и визуальное подтверждение:** мы количественно оценим экономию вычислительных ресурсов в нашем подходе и предоставим качественные доказательства с помощью визуализации t-SNE с целью подтверждения разделимости классов извлеченных вложений, отражает важность трансферного обучения и позволяет создавать интерпретируемые визуализации признаков..

2. СВЯЗАННЫЕ РАБОТЫ

Рассмотрим ландшафт глубокого обучения в эндоскопии ЖКТ, от доминирующей парадигмы полного дообучения до более эффективных стратегий трансферного обучения.

2.1. Полное дообучение в эндоскопии

Доминирующей парадигмой в анализе эндоскопических изображений является полное дообучение, при котором все параметры модели, предварительно обученной на общем наборе данных ImageNet обновляются на целевых медицинских данных для достижения самых современных (SOTA) результатов [13]. На бенчмарк-наборах данных Kvasir [14] и HyperKvasir [15] полностью дообученные CHC, такие как DenseNet-201 [16] и ResNet-101, продемонстрировали точность классификации, превышающую 95–97%. В последнее время Vision Transformers (ViT) в задачах эндоскопической классификации достигли даже лучших результатов, чем CHC [17]. Однако это сопряжено с большими вычислительными затратами, что является значительным препятствием для быстрого экспериментирования и клинического внедрения [18].

2.2. Эффективное трансферное обучение и извлечение признаков

Чтобы смягчить вычислительную нагрузку полного дообучения, были разработаны более эффективные стратегии. Методы параметроэффективного дообучения (PEFT), такие как LoRA [19] или техники разреживания [20], обновляют лишь небольшую долю параметров модели, что позволяет снизить затраты на обучение. Другим подходом является извлечение признаков, при котором весь предварительно обученный кодировщик замораживается и обучается только простой классификационный декодер на признаках, которые он производит. Этот метод,

также известный как линейное зондирование (linear probing), когда используется линейный слой [21], сокращает время обучения с часов до минут.

Однако извлечение признаков в медицинской визуализации обычно основывалось на эвристике использования только последнего слоя кодировщика. Этот подход упускает богатую, специфичную для задачи информацию, доступную в промежуточных слоях [22, 23]. Насколько нам известно, систематический послойный анализ для определения оптимального источника признаков для эндоскопической классификации не проводился, он и рассматривается в настоящей работе.

3. МЕТОДОЛОГИЯ

Наша методология использует простой конвейер для выделения качества характеристик предварительно обученных признаков в качестве основной переменной. Наша экспериментальная структура включает замороженную основу Vision Transformer (ViT) для генерации признаков, процесс послойного извлечения и неглубокую обучаемую классификационную надстройку.

3.1. Архитектурный конвейер

Предлагаемая архитектура показана на рис. 1. Входное эндоскопическое изображение сначала проходит через предварительно обученный и полностью замороженный кодировщик ViT. Мы можем формально определить кодировщик Φ как композицию из N блоков трансформера, $\Phi = L_N \circ \ldots \circ L_1$. Затем мы перехватываем выходную карту признаков из определенного промежуточного слоя L_i, где $i \in \{1,2, \ldots ,24\}$. Эта высокоразмерная карта признаков $\mathbf{z}_i(x) = (L_i \circ \ldots \circ L_1)(x)$ агрегируется в единый вектор признаков $\mathbf{h}_i(x)$, который служит входом для неглубокого обучаемого декодера, ответственного за конечное предсказание класса. Конечное распределение вероятностей $\hat{y}$ задается формулой

$$\hat{y} = \text{softmax}(\mathbf{W}\mathbf{h}_i(x) + \mathbf{b}),$$

где $\{\mathbf{W}, \mathbf{b}\}$ – единственные обучаемые параметры модели. Весь этот процесс повторяется независимо для каждого из 24 слоев кодировщика.

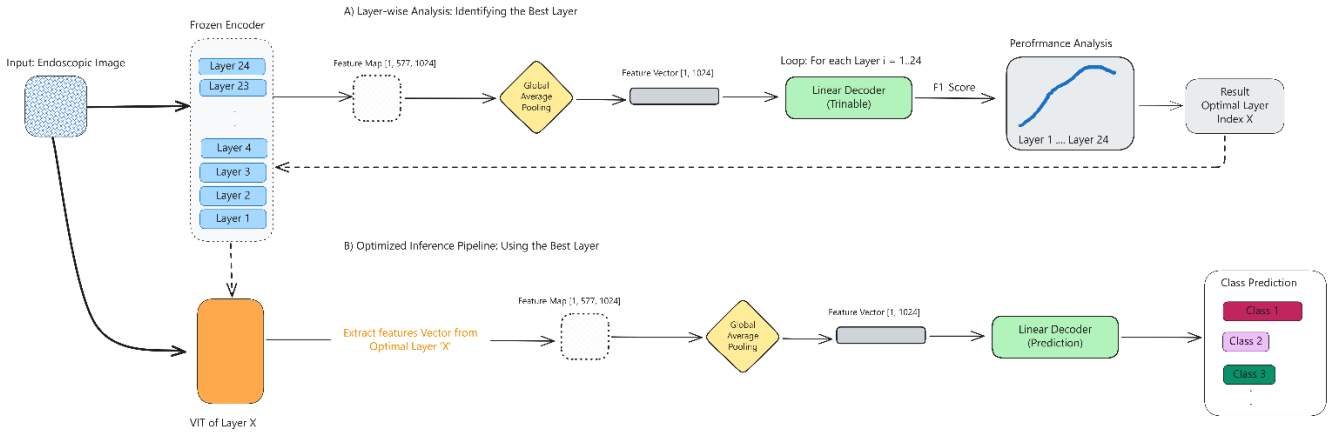

Рис. 1. Обзор предлагаемого конвейера классификации.

Примечание. Входное изображение обрабатывается замороженным кодировщиком Vision Transformer (ViT). Мы извлекаем вложения токенов патчей из определенного промежуточного слоя *X*. Эти вложения агрегируются с помощью глобального среднего пулинга для формирования единого вектора признаков, который затем подается в минималистичный линейный декодер для окончательной классификации. Декодер является единственным обучаемым компонентом в архитектуре.

3.2. Наборы данных и предварительная обработка

Для обеспечения надежности и обобщаемости полученных результатов мы проверили верификацию предложенного метода на трех широко известных публичных наборах данных: Kvasir-V1 и Kvasir-V2 [14], которые являются сбалансированными наборами данных, содержащими 4000 и 8000 изображений соответственно, по 8 классам находок в ЖКТ, и HyperKvasir [15], крупномасштабный набор данных, из которого мы формируем задачу классификации на 8 классах из 8531 изображения, чтобы соответствовать нашим другим экспериментам. Для всех экспериментов мы разделяем данные на обучающий (80%) и валидационный (20%) наборы, используя стратифицированную выборку для сохранения распределения классов в обеих частях. Все изображения изменяются до нативного разрешения кодировщика 336× 336 пикселей и нормализуются с использованием стандартного среднего значения и стандартного отклонения ImageNet [10].

3.3. Замороженный кодировщик и послойное извлечение признаков

$$\mathcal{L}_{\text{pretrain}} = \frac{1}{2B}\sum_{i=1}^{B}\left(\mathcal{L}_i^{(v\to u)} + \mathcal{L}_i^{(u\to v)}\right),$$

где

$$\mathcal{L}_i^{(v\to u)} = -\log\frac{\exp\left(\frac{\text{sim}(\mathbf{v}_i,\mathbf{u}_i)}{\tau}\right)}{\sum_{j=1}^{B}\exp\left(\frac{\text{sim}(\mathbf{v}_i,\mathbf{u}_j)}{\tau}\right)},$$

$$\mathcal{L}_i^{(u\to v)} = -\log\frac{\exp\left(\frac{\text{sim}(\mathbf{u}_i,\mathbf{v}_i)}{\tau}\right)}{\sum_{j=1}^{B}\exp\left(\frac{\text{sim}(\mathbf{u}_i,\mathbf{v}_j)}{\tau}\right)}.$$

Здесь $\mathcal{L}_i^{(v\to u)}$ – потери от изображения к тексту, которые приближают вложение изображения $\mathbf{v}_i$ к соответствующему текстовому вложению $\mathbf{u}_i$; $\mathcal{L}_i^{(u\to v)}$ – потери от текста к изображению, которые обеспечивают близость текстового вложения $\mathbf{u}_i$ к парному изображению $\mathbf{v}_i$. Это было сделано для создания надежного общего пространства вложений.

Для проведения послойного анализа мы регистрируем прямой хук (forward hook) на каждом из 24 остаточных блоков ViT. Для такого входного изображения выход блока на слое i представляет собой тензор формы [1, 577, 1024], соответственно представляющий размер батча, длину последовательности и размерность признаков. Длина последовательности 577 состоит из одного токена [CLS] и 576 токенов патчей (24×24).

Для принципиального сравнения по всем слоям мы отбрасываем специализированный токен [CLS], репрезентативная система которого используется для обработки последнего слоя и не является однородной по всей глубине сети. Вместо этого мы сосредотачиваемся на сетке из 576 токенов патчей, которая представляет пространственную карту признаков изображения на любом данном слое i. Мы агрегируем эти вложения патчей с помощью глобального среднего пулинга

[24] для получения единого, концептуально последовательного вектора признаков, что позволяет провести справедливую оценку основных визуальных признаков на каждой глубине.

3.4. Легковесный классификационный декодер

Мы используем простой декодер: один полносвязный линейный слой без скрытых слоев или нелинейных активаций. Он напрямую отображает 1024-мерный вектор признаков из кодировщика в N_{classes} выходных логитов для классификации.

Этот выбор намеренно минимизирует количество обучаемых параметров, что изолирует вклад в производительность замороженных признаков. Для наших экспериментов с 8 классами этот декодер содержит всего 1024×8=8192 обучаемых веса, что на несколько порядков меньше по сравнению с миллионами параметров при полном дообучении. Это также важно для снижения риска переобучения на небольших медицинских наборах данных и ускорения процесса обучения.

3.5. Экспериментальный протокол и оценка

Для каждого из 24 слоев мы обучаем наш линейный декодер с нуля со случайной инициализацией. Модель обучается в течение 30 эпох с использованием оптимизатора Adam [25] с скоростью обучения 1×10^{-4} и размером батча 8. В качестве целевой функции используем стандартную кросс-энтропийную функцию потерь.

Далее оцениваем признаки каждого слоя с помощью набора стандартных метрик классификации: точность (Accuracy), F1-мера (Macro), точность (Precision) и полнота (Recall). Качество признаков на данном слое i можно формально определить через минимальную *ошибку линейного зондирования* $\mathcal{E}_i$, достижимую оптимальным линейным классификатором:

$$\mathcal{E}_i(\mathcal{D}) = \min_{\mathbf{W},\mathbf{b}} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{I}(\text{argmax}(\mathbf{W}\mathbf{h}_i(x) + \mathbf{b}) \neq y)],$$

где $\mathbb{I}(\cdot)$ — индикаторная функция. Наши метрики служат эмпирическими оценками $1 - \mathcal{E}_i(\mathcal{D}_{\text{val}})$. Построив их график в зависимости от индекса слоя, мы можем

определить слой, который дает лучшие признаки для классификации эндоскопических изображений.

4. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

В этом разделе мы представляем результаты исследования. Сначала мы хотим оценить эффективность предложенного кодировщика, сравнивая его оптимальную производительность с самыми современными бенчмарками. Затем мы представляем послойный анализ качества признаков. Наконец, мы предоставляем качественную валидацию через визуализации пространства вложений и анализ поведения классификации оптимального слоя.

4.1. Общая производительность по сравнению с современным уровнем

Чтобы проверить наш конвейер, мы сначала определили наиболее производительный слой из нашего анализа (подробно см. в разд. 4.2) и сравнили эту оптимальную конфигурацию с несколькими устоявшимися, полностью дообученными моделями.

Табл. 1 и 2 представляют это сравнение. Наш метод, обозначенный как PE-Core-L21 + Linear, использует признаки, извлеченные из 21-го слоя замороженного кодировщика PE-Core, которые подаются в простой линейный классификатор.

Табл. 1. Результаты бенчмаркинга на наборе данных Kvasir v1. Лучшие результаты выделены жирным шрифтом. N/P указывает, что метрика не была предоставлена в исходной статье.

Метод	Точн.	Precision	Recall	F1-Score	Обуч. парам. (M)
Deep Ensemble [26]	**98.45**	**0.99**	**0.97**	**0.97**	10.3
ResNet 50 [8]	96.57	N/P	N/P	N/P	23.8
NasNet-Mobile [27]	94.53	N/P	N/P	N/P	5.3
PE Core (L21) + Linear	92.37	0.9251	0.9237	0.9236	**0.008**
EfficientNet [28]	92.28	N/P	N/P	N/P	5.3
Inception V3 [29]	91.57	N/P	N/P	N/P	25.6

Табл. 2. Результаты бенчмаркинга на наборе данных Kvasir v2. Лучшие результаты выделены жирным шрифтом. N/P указывает, что метрика не была предоставлена в исходной статье.

Метод	Точн.	Precision	Recall	F1-Score	Обуч. парам. (M)
Deep Ensemble [26]	**97.83**	**0.98**	**0.97**	**0.96**	10.3
NasNet-Mobile [27]	93.21	N/P	N/P	N/P	5.3
PE Core (L21) + Linear	92.63	0.9298	0.9262	0.9256	**0.008**
ResNet 50 [8]	90.58	N/P	N/P	N/P	23.8
EfficientNet [28]	90.28	N/P	N/P	N/P	5.3
Inception V3 [29]	88.38	N/P	N/P	N/P	25.6

Для дальнейшей проверки обобщающих способностей нашего подхода мы расширили оценку на набор данных HyperKvasir. В этой задаче наша модель с использованием признаком из оптимального 21-го слоя, достигла точности 93.4% и Macro F1-Score 93.12%. Такая высокая производительность, полученная без какого-либо дообучения и с аналогичными конфигурациями обучения, свидетельствует о том, что признаки из замороженного кодировщика не только эффективны на сбалансированных данных, но и достаточно надежны для хорошего обобщения на разных наборах данных.

4.2. Послойный анализ признаков: определение оптимальной глубины

Согласно [30], лучшие признаки не всегда выявляются в последнем слое. На рис. 2 показаны четыре метрики в зависимости от индекса слоя кодировщика для набора данных Kvasir-V2.

Наш послойный анализ показал ясную и последовательную картину для всех трех наборов данных. Как и можно было ожидать, производительность в самых неглубоких слоях (например, 1–5) была низкой, особенно на 8-классовых наборах данных Kvasir, вероятно, из-за сосредоточенности на общих, низкоуровневых признаках, таких как края и цвета. Мы наблюдали значительный рост производительности по мере продвижения к средним слоям (приблизительно 6–18), так как модель переходит от простых паттернов к абстрактным, семантически богатым представлениям, которые важны для дифференцирования сложных эндоскопических патологий.

Наши результаты последовательно показывают, что наиболее отличительные признаки находятся в позднестадийных слоях. Как видно на рис. 2, производительность неуклонно растет, достигая пика на 21-м слое. После этой точки наблюдается небольшое ухудшение признаков в последних одном или двух слоях. Этот феномен можно объяснить с точки зрения принципа информационного бутылочного горлышка [31], где взаимная информация между признаками $\mathbf{Z}_i$ и последующей задачей Y_{down} максимизируется на промежуточном слое i^*, в то время как информация, относящаяся к задаче предварительного обучения Y_{pre}, продолжает уточняться:

$$i^* = \arg\max_{i \in \{1,\ldots,N\}} I(\mathbf{Z}_i; Y_{\text{down}}) \quad \text{s.t.} \quad \frac{\partial}{\partial i} I\left(\mathbf{Z}_i; Y_{\text{pre}}\right) \geq 0.$$

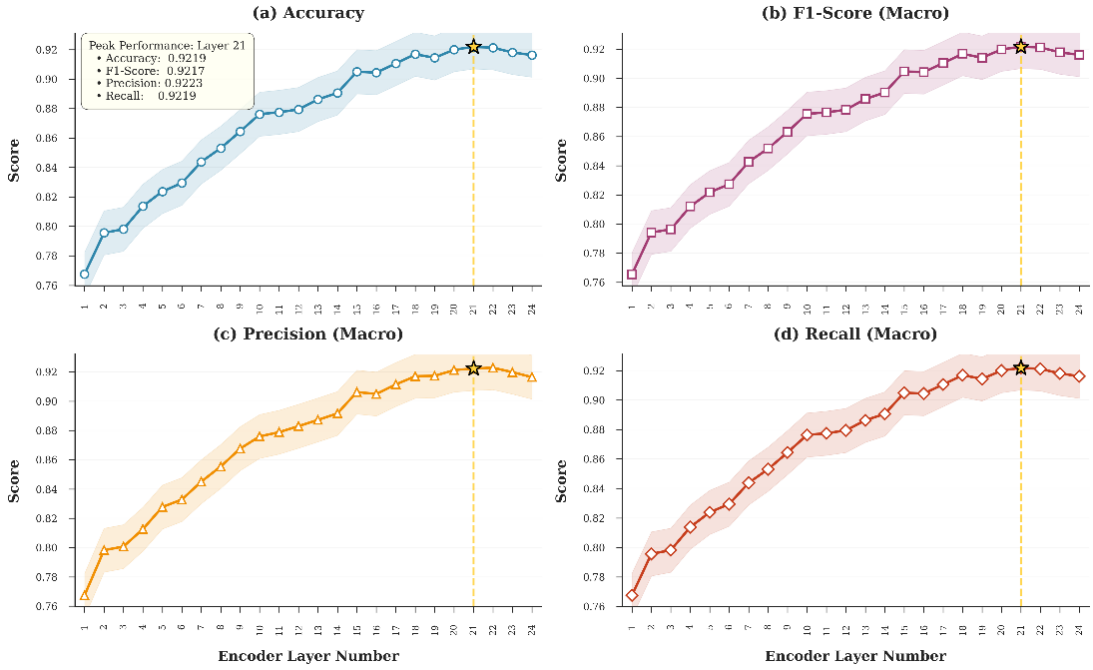

Рис. 2. Послойная производительность на наборе данных Kvasir-V2. Четыре метрики (ось *Y*) нанесены на график в зависимости от индекса слоя кодировщика (ось *X*). Производительность резко возрастает от неглубоких к среднеглубоким слоям, достигая пика на слое 21, а затем немного снижается на последних слоях.

Этот паттерн указывает на место, где признаки достигли максимальной семантической сложности для нашей задачи классификации и непосредственно перед тем, как стать специализированными для исходной цели предварительного

обучения кодировщика. На основе проведенного анализа, мы выбрали 21-й слой в качестве оптимального источника признаков для наших сравнений.

4.3. Качественная валидация качества признаков

Чтобы обеспечить интуитивное понимание полученных количественных результатов, мы провели два качественных анализа признаков, извлеченных из нашего оптимального слоя.

4.3.1. Визуализация пространства вложений

Чтобы дать интуитивное, качественное понимание того, почему признаки из определенного нами оптимального слоя так эффективны, мы визуализируем их структуру в 2D-пространстве. На рис. 3 представлены 1024-мерные вложения признаков из 21-го слоя, спроецированные с использованием алгоритма снижения размерности t-SNE [32].

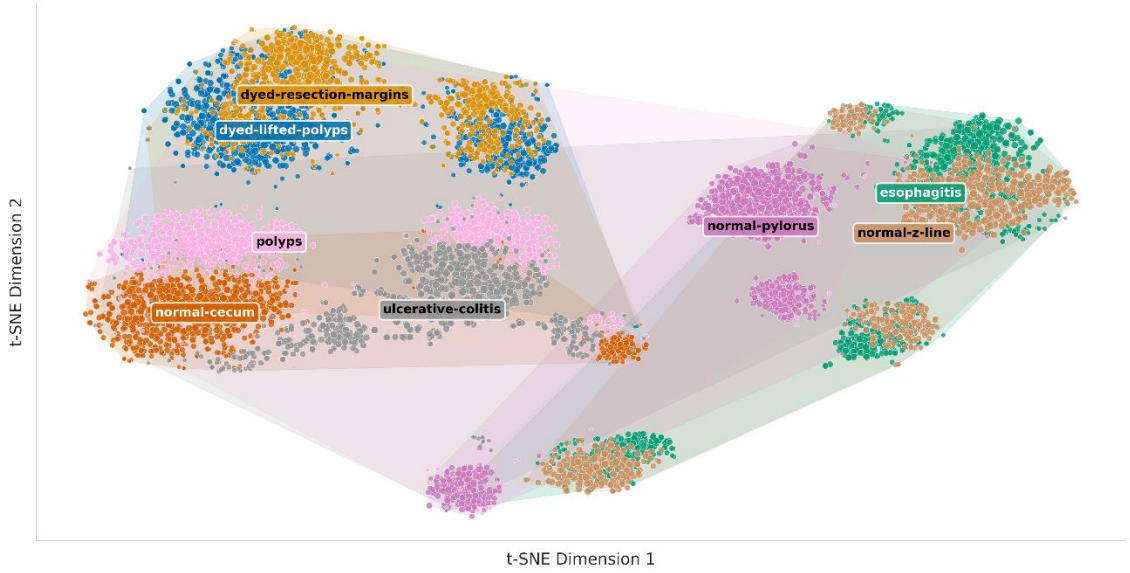

Рис. 3. Проекция t-SNE вложений признаков из оптимального 21-го слоя на наборе данных Kvasir-V2. Каждый цвет представляет отдельный класс. Проекция показывает хорошо разделенные кластеры благодаря высокоразличимому и линейно разделимому пространству признаков.

Результаты обоих методов согласуются и являются визуально убедительными. Вложения образуют плотные, хорошо разделенные кластеры, причем каждый кластер соответствует отдельному эндоскопическому классу. Такая высокая

степень линейной разделимости в пространстве признаков напрямую подтверждает наши количественные выводы.

4.3.2. Анализ структуры изученного пространства признаков

Помимо высоких количественных метрик, анализ пространства признаков 21-го слоя выявил структуру, которая соответствует клиническому состоянию тканей. Пространство признаков не произвольно кластеризовано, а имеет осмысленную проекцию, отражающую клиническое сходство. Например, классы, связанные процедурными артефактами, такие как окрашенные приподнятые полипы и окрашенные края резекции, образуют отдельные, но смежные кластеры. Аналогично, анатомически связанные классы, такие как нормальная *Z*-линия и эзофагит, группируются рядом друг с другом, что важно, поскольку эзофагит — это воспаление, возникающее на *Z*-линии. Особенно показательная группировка происходит с классами полипов, язвенного колита и нормальной слепой кишки, которые имеют схожую подстилающую текстуру слизистой оболочки. Мы можем количественно оценить эту группировку, измерив внутриклассовую дисперсию карт пространственного соответствия:

$$\sigma^2_{\text{spatial}}(i,c) = \frac{1}{M^2}\sum_{j,k} \text{Var}_{x\in\mathcal{D}_c}\left(\mathbf{S}_i(x)_{jk}\right),$$

где $\mathbf{S}_i(x)$ — матрица попарных косинусных сходств между токенами патчей. Тот факт, что эти классы попадают в одну и ту же общую область, говорит о том, что вложение достаточно сильное, чтобы идентифицировать внешний вид этой ткани на основе общих визуальных характеристик. Этот общий семантический анализ демонстрирует, что замороженный кодировщик действует не просто как распознаватель образов, а как сложный экстрактор признаков, который захватывает иерархию визуальной информации — от процедурных артефактов до анатомического контекста, что оправдывает его высокую производительность.

5. ОБСУЖДЕНИЕ

Проведенное эмпирическое исследование дает ответы, имеющие важное значение для будущего развития и внедрения медицинских систем ИИ.

5.1. Основные выводы: проверка основных гипотез

Наши результаты напрямую подтверждают основную гипотезу: замороженный кодировщик с линейным декодером достигает конкурентоспособной производительности на нескольких эндоскопических бенчмарках. Это было достигнуто всего с 8000 обучаемыми параметрами и минимальным обучением, при этом потери на валидации все еще имели тенденцию к снижению на момент завершения, что раскрывает мощь предварительно обученных признаков и демонстрирует, что обширное дообучение не является обязательным условием для высококачественной классификации медицинских изображений.

Центральным для нашего исследования является то, что систематический анализ подтвердил гипотезу о том, что оптимальные признаки для последующей задачи не всегда находятся в последнем слое. Мы эмпирически определили «золотую середину» в позднепромежуточных слоях, в частности, в 21-м слое, где производительность классификации была самой высокой перед небольшим падением в последних слоях (рис. 2). Падение производительности в последних слоях можно объяснить целью предварительного обучения кодировщика. Последние слои оптимизированы для исходной задачи «зрение — язык», сжимая пространство признаков и отбрасывая тонкую визуальную информацию, которая важна для медицинской диагностики. Поздне-промежуточный слой, такой как 21-й слой, предоставляет полные семантические знания без этого сжатия, специфического для задачи. На этой глубине сеть понимает сложные медицинские концепции, такие как текстура ткани, что согласуется с принципом «информационного бутылочного горлышка», где итоговое обобщение может быть слишком агрессивным в неучете деталей, важных для новой задачи. Поэтому лучшие признаки не обязательно находятся в последнем слое.

5.2. Анализ поведения модели и асимметричные ошибки

Полученные результаты демонстрируют клинически значимую асимметрию между нормальной *Z*-линией и эзофагитом. Модель неверно классифицирует изображение нормальной *Z*-линии как эзофагит в 23.5% случаев, тогда как обратная ошибка имеет место только в 4.5% случаев. Это не случайный сбой. *Z*-линия — это место, где возникает эзофагит, и ранние или легкие случаи могут быть визуально похожи на нормальную *Z*-линию [33]. Поэтому модель, обученная быть чувствительной к патологическим признакам, классифицирует пограничнонормальную *Z*-линию как эзофагит. Явный эзофагит имеет признаки, отсутствующие на нормальной *Z*-линии, отсюда и более низкая обратная ошибка.

ЗАКЛЮЧЕНИЕ

Настоящее исследование отвечает на простой, но фундаментальный вопрос: всегда ли мы должны выполнять дообучение для достижения высокой производительности? Ответ: нет. Мы продемонстрировали, что предварительно обученный фиксированный кодировщик с признаками из оптимальной глубины обеспечивает мощную и эффективную основу для классификации эндоскопических изображений. Основной вклад заключался в том, чтобы отобразить качество признаков слой за слоем, предполагая, что лучшие представления существуют непосредственно перед тем, как модель становится чрезмерно специализированной. Полученные результаты имеют важное значение для практических применений и позволят в будущем создавать более простые и быстрые системы ИИ, подходящие для реальной клинической практики.

Благодарности

Работа поддержана Академией наук Республики Татарстан в рамках грантового соглашения № 254/2024-ПД.

СПИСОК ЛИТЕРАТУРЫ

1. *Abusuliman M., Jamali T., Zuchelli T.E.* Advances in gastrointestinal endoscopy: A comprehensive review of innovations in cancer diagnosis and management // World Journal of Gastrointestinal Endoscopy. 2025. Vol. 17, No. 5. P. 105468.

2. *Simadibrata D.M., Lesmana E., Fass R.* Role of endoscopy in gastroesophageal reflux disease // Clinical Endoscopy. 2023. Vol. 56, No. 6. P. 681–692.

3. *Mathews A.A., Draganov P.V., Yang D.* Endoscopic management of colorectal polyps: From benign to malignant polyps // World Journal of Gastrointestinal Endoscopy. 2021. Vol. 13, No. 9. P. 356.

4. *Bernatchi I.N., Voidazan S., Petrut M.I., Gabos G., Balasescu M., Nicolau C.* Inter-observer variability on the value of endoscopic images for the documentation of upper gastrointestinal endoscopy – our center experience // Acta Marisiensis – Seria Medica. 2023.

5. *Ghazi G.G.R.J.J. et al.* Sampling error in the diagnosis of colorectal cancer is associated with delay to surgery: a retrospective cohort study // Surgical Endoscopy. 2022. Vol. 36. P. 4893–4902.

6. *Khalifa M., Albadawy M.* Ai in diagnostic imaging: Revolutionising accuracy and efficiency // Computer Methods and Programs in Biomedicine Update. 2024. Vol. 5. P. 100146.

7. *LeCun Y., Bengio Y., Hinton G.* Deep learning // Nature. 2015. Vol. 521, No. 7553. P. 436–444.

8. *He K., Zhang X., Ren S., Sun J.* Deep residual learning for image recognition // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015. P. 770-778.

9. *Dosovitskiy A. et al.* An image is worth 16x16 words: Transformers for image recognition at scale // 3rd Conference on Neural Information Processing Systems. 2021.

10. *Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L.* Imagenet: A large-scale hierarchical image database // 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009. P. 248–255.

11. *Su S.S. et al.* Democratizing protein language models with parameter-efficient fine-tuning // Proceedings of the National Academy of Sciences of the United States of America. 2024. Vol. 121. P. e2405840121.

12. *Sanchez-V T.S., Rahimi A., Oktay O., Bharadwaj S.* Addressing the exorbitant cost of labeling medical images with active learning // International Conference on Machine Learning and Medical Imaging Analysis.

13. *Zhang Z.Z. et al.* Active, continual fine tuning of convolutional neural networks for reducing annotation efforts // Medical Image Analysis. 2021. Vol. 71. P. 101997.

14. *Pogorelov K. et al.* Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection // Proceedings of the 8th ACM on Multimedia Systems Conference. 2017. P. 164–169.

15. *Borgli H. et al.* Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy // Scientific Data. 2020. Vol. 7, No. 1. P. 283.

16. *Huang G., Liu Z., van der Maaten L., Weinberger K.Q.* Densely connected convolutional networks // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018. P. 2261–2269.

17. *Shah S.T. et al.* Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review // Journal of Medical Systems. 2024. Vol. 48, No. 1. P. 84.

18. *Rosenthal J.T., Beecy A., Sabuncu M.R.* Rethinking clinical trials for medical ai with dynamic deployments of adaptive systems // npj Digital Medicine. 2025. Vol. 8, No. 1. P. 252.

19. *Hu E.J. et al.* Lora: Low-rank adaptation of large language models // International Conference on Learning Representations. 2021.

20. *Farina M., Ahmad U., Taha A., Younes H., Mesbah Y., Yu X., Pedrycz W.* Sparsity in transformers: A systematic literature review // Neurocomputing. 2024. Vol. 582. P. 127468.

21. *Chen T., Kornblith S., Norouzi M., Hinton G.* A simple framework for contrastive learning of visual representations // Proceedings of the 37th International Conference on Machine Learning. 2020. P. 1597–1607.

22. *Yan Y.C. et al.* Brain tumor intelligent diagnosis based on auto-encoder and u-net feature extraction // PLOS ONE. 2025. Vol. 20, No. 3. P. e0315631.

23. *Jawahar B.S.G., Seddah D.* What does bert learn about the structure of language? // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. P. 3651–3657.

24. *Lin M., Chen Q., Yan S.* Network in network // 2nd International Conference on Learning Representations, ICLR 2014. 2014.

25. *Kingma D.P., Ba J.* Adam: A method for stochastic optimization // 5th International Conference on Learning Representations, ICLR 2017. 2017.

26. *Siddiqui S., Khan J.A., Algamdi S.* Deep ensemble learning for gastrointestinal diagnosis using endoscopic image classification // PeerJ Computer Science. 2025. Vol. 11. P. e2809.

27. *Zoph B., Vasudevan V., Shlens J., Le Q.V.* Learning transferable architectures for scalable image recognition // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018. P. 8697-8705.

28. *Tan M., Le Q.V.* Efficientnet: Rethinking model scaling for convolutional neural networks // Proceedings of the 36th International Conference on Machine Learning. 2020. P. 6105–6114.

29. *Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z.* Rethinking the inception architecture for computer vision // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015. P. 2818–2826.

30. *Ben-Younes D. et al.* Perception encoder: The best visual embeddings are not at the output of the network // The Twelfth International Conference on Learning Representations. 2025.

31. *Tishby N., Pereira F.C., Bialek W.* The information bottleneck method // 37th Annual Allerton Conference on Communication, Control, and Computing. 2000.

32. van der Maaten L., Hinton G. Visualizing data using t-sne // Journal of Machine Learning Research. 2008. Vol. 9. P. 2579–2605.

33. *Kamboj A.K., Gaddam S., Lo S.K., Rezaie A.* Irregular z-line: To biopsy or not to biopsy? // Digestive Diseases and Sciences. 2024. Vol. 69, No. 8. P. 2734–2740.

WHERE DO THE BEST FEATURES LIE? A LAYER-WISE ANALYSIS OF FROZEN ENCODERS FOR EFFICIENT ENDOSCOPIC IMAGE CLASSIFICATION

A. Taha[1] [0009-0006-6346-4162], R. A. Lukmanov[2] [0000-0001-9257-7410]

[1, 2]Innopolis University, Innopolis, Russia

[1]Center of Artificial Intelligence at Innopolis University, Innopolis, Russia

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

Abstract

In our quest to advance medical AI, we demonstrate that a pre-trained and frozen Vision Transformer paired with a linear classifier can achieve highly competitive performance in endoscopic image classification. Our central contribution is a systematic, layer-wise analysis that identifies the source of the most powerful features, challenging the common heuristic of using only the final layer. We uncover a distinct "peak-before-the-end" phenomenon, where a late-intermediate layer offers a more generalizable representation for the downstream medical task. On the Kvasir and HyperKvasir benchmarks, our parameter-light approach not only achieves excellent accuracy but also drastically reduces computational overhead. This work provides a practical roadmap for efficiently leveraging the power of general foundation models in clinical environments.

Keywords*: endoscopic image classification, frozen encoder, feature extraction, layer-wise analysis, vision transformer (ViT), transfer learning, self-supervised learning (SSL), medical AI.*

REFERENCES

1. *Abusuliman M., Jamali T., Zuchelli T.E.* Advances in gastrointestinal endoscopy: A comprehensive review of innovations in cancer diagnosis and management // World Journal of Gastrointestinal Endoscopy. 2025. Vol. 17, No. 5. P. 105468.

2. *Simadibrata D.M., Lesmana E., Fass R.* Role of endoscopy in gastroesophageal reflux disease // Clinical Endoscopy. 2023. Vol. 56, No. 6. P. 681–692.

3. *Mathews A.A., Draganov P.V., Yang D.* Endoscopic management of colorectal polyps: From benign to malignant polyps // World Journal of Gastrointestinal Endoscopy. 2021. Vol. 13, No. 9. P. 356.

4. *Bernatchi I.N., Voidazan S., Petrut M.I., Gabos G., Balasescu M., Nicolau C.* Inter-observer variability on the value of endoscopic images for the documentation of upper gastrointestinal endoscopy – our center experience // Acta Marisiensis – Seria Medica. 2023.

5. *Ghazi G.G.R.J.J. et al.* Sampling error in the diagnosis of colorectal cancer is associated with delay to surgery: a retrospective cohort study // Surgical Endoscopy. 2022. Vol. 36. P. 4893–4902.

6. *Khalifa M., Albadawy M.* Ai in diagnostic imaging: Revolutionising accuracy and efficiency // Computer Methods and Programs in Biomedicine Update. 2024. Vol. 5. P. 100146.

7. *LeCun Y., Bengio Y., Hinton G.* Deep learning // Nature. 2015. Vol. 521, No. 7553. P. 436–444.

8. *He K., Zhang X., Ren S., Sun J.* Deep residual learning for image recognition // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015. P. 770-778.

9. *Dosovitskiy A. et al.* An image is worth 16x16 words: Transformers for image recognition at scale // 3rd Conference on Neural Information Processing Systems. 2021.

10. *Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L.* Imagenet: A large-scale hierarchical image database // 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009. P. 248–255.

11. *Su S.S. et al.* Democratizing protein language models with parameter-efficient fine-tuning // Proceedings of the National Academy of Sciences of the United States of America. 2024. Vol. 121. P. e2405840121.

12. *Sanchez-V T.S., Rahimi A., Oktay O., Bharadwaj S.* Addressing the exorbitant cost of labeling medical images with active learning // International Conference on Machine Learning and Medical Imaging Analysis.

13. *Zhang Z.Z. et al.* Active, continual fine tuning of convolutional neural networks for reducing annotation efforts // Medical Image Analysis. 2021. Vol. 71. P. 101997.

14. *Pogorelov K. et al.* Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection // Proceedings of the 8th ACM on Multimedia Systems Conference. 2017. P. 164–169.

15. *Borgli H. et al.* Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy // Scientific Data. 2020. Vol. 7, No. 1. P. 283.

16. *Huang G., Liu Z., van der Maaten L., Weinberger K.Q.* Densely connected convolutional networks // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018. P. 2261–2269.

17. *Shah S.T. et al.* Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review // Journal of Medical Systems. 2024. Vol. 48, No. 1. P. 84.

18. *Rosenthal J.T., Beecy A., Sabuncu M.R.* Rethinking clinical trials for medical ai with dynamic deployments of adaptive systems // npj Digital Medicine. 2025. Vol. 8, No. 1. P. 252.

19. *Hu E.J. et al.* Lora: Low-rank adaptation of large language models // International Conference on Learning Representations. 2021.

20. *Farina M., Ahmad U., Taha A., Younes H., Mesbah Y., Yu X., Pedrycz W.* Sparsity in transformers: A systematic literature review // Neurocomputing. 2024. Vol. 582. P. 127468.

21. *Chen T., Kornblith S., Norouzi M., Hinton G.* A simple framework for contrastive learning of visual representations // Proceedings of the 37th International Conference on Machine Learning. 2020. P. 1597–1607.

22. *Yan Y.C. et al.* Brain tumor intelligent diagnosis based on auto-encoder and u-net feature extraction // PLOS ONE. 2025. Vol. 20, No. 3. P. e0315631.

23. *Jawahar B.S.G., Seddah D.* What does bert learn about the structure of language? // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. P. 3651–3657.

24. *Lin M., Chen Q., Yan S.* Network in network // 2nd International Conference on Learning Representations, ICLR 2014. 2014.

25. *Kingma D.P., Ba J.* Adam: A method for stochastic optimization // 5th International Conference on Learning Representations, ICLR 2017. 2017.

26. *Siddiqui S., Khan J.A., Algamdi S.* Deep ensemble learning for gastrointestinal diagnosis using endoscopic image classification // PeerJ Computer Science. 2025. Vol. 11. P. e2809.

27. *Zoph B., Vasudevan V., Shlens J., Le Q.V.* Learning transferable architectures for scalable image recognition // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018. P. 8697-8705.

28. *Tan M., Le Q.V.* Efficientnet: Rethinking model scaling for convolutional neural networks // Proceedings of the 36th International Conference on Machine Learning. 2020. P. 6105–6114.

29. *Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z.* Rethinking the inception architecture for computer vision // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015. P. 2818–2826.

30. *Ben-Younes D. et al.* Perception encoder: The best visual embeddings are not at the output of the network // The Twelfth International Conference on Learning Representations. 2025.

31. *Tishby N., Pereira F.C., Bialek W.* The information bottleneck method // 37th Annual Allerton Conference on Communication, Control, and Computing. 2000.

32. van der Maaten L., Hinton G. Visualizing data using t-sne // Journal of Machine Learning Research. 2008. Vol. 9. P. 2579–2605.

33. *Kamboj A.K., Gaddam S., Lo S.K., Rezaie A.* Irregular z-line: To biopsy or not to biopsy? // Digestive Diseases and Sciences. 2024. Vol. 69, No. 8. P. 2734–2740.

СВЕДЕНИЯ ОБ АВТОРАХ

Ахмад ТАХА — 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 Ph.D. reseacher and a researcher at the Center of Artificial Intelligence at 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.

Research interests: Medical AI, Self-Supervised Learning (SSL), Transformers, Natural Language Processing (NLP), Computer Vision, Machine Learning.

email: a.taha@innopolis.university

ORCID: 0009-0006-6346-4162

Рустам А. ЛУКМАНОВ *(*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 A. LUKMANOV (PhD, University of Bern, 2021) is a researcher and associate professor specializing in machine learning, bioinformatics, data analysis, and explainable AI. He is a recipient of the BRICS and SCO Young Leaders Award (2023). He teaches courses on explainable AI and knowledge representation at Innopolis University.

email: r.lukmanov@innopolis.university

ORCID: 0000-0001-9257-7410

Материал поступил в редакцию 1 октября 2025 года