DOI 10.37882/2223-2966.2025.01-2.09

ИЕРАРХИЧЕСКИЕ СЕТИ ТИПА ТРАНСФОРМЕР ДЛЯ ОБНАРУЖЕНИЯ АНОМАЛИЙ НА ВИДЕОРЯДАХ С КАМЕР ВИДЕОНАБЛЮДЕНИЯ

HIERARCHIAL TRANSFORMER NETWORKS FOR ANOMALY DETECTION IN SURVEILLANCE VIDEOS

A. Gultiaev

Summary. This paper presents a novel approach for anomaly detection in surveillance videos using hierarchical transformer networks without relying on convolutional neural networks. We leverage Video Vision Transformers (ViViT) combined with contrastive learning to extract meaningful embeddings from video segments. To handle variable-length video clips, we introduce a hierarchical transformer architecture that captures both segment-level and event-level representations. Trained on the DCSASS dataset, our method demonstrates significant improvements in classification, clustering, and anomaly detection tasks compared to traditional approaches. Our results indicate that the proposed model can effectively assist surveillance operators in detecting abnormal activities, thereby enhancing security measures.

Keywords: machine learning, artificial intelligence, computer vision, neural network, transformer, contrastive learning, vector embedding, classification, clustering, anomaly detection.

Гультяев Андрей Андреевич
Аспирант, Национальный Исследовательский Ядерный Университет «МИФИ»
angultiaev@gmail.com

Аннотация. В статье представлен новый подход к обнаружению аномалий на видеозаписях с камер видеонаблюдения с использованием иерархических нейронных сетей типа Transformer без использования сверточных нейронных сетей. Для извлечения векторных представлений из видеосегментов использована архитектура Video Vision Transformer (ViViT) в сочетании с подходом к обучению, называемым контрастным обучением. Для работы с видеозаписями переменной длины введена иерархическая архитектура сетей Transformer, которая получает представления как на уровне сегментов видео, так и на уровне событий. Обученный на наборе данных DCSASS, метод демонстрирует значительное улучшение в задачах классификации, кластеризации и обнаружения аномалий по сравнению с традиционными подходами. Результаты показывают, что предложенная модель может эффективно помочь операторам видеонаблюдения в обнаружении аномальных действий, тем самым повышая эффективность мер безопасности.

Ключевые слова: машинное обучение, искусственный интеллект, компьютерное зрение, нейронная сеть, трансформер, контрастное обучение, векторное представление, классификация, кластеризация, обнаружение аномалий.

Введение

Системы видеонаблюдения играют важнейшую роль в обеспечении безопасности в общественных местах. С увеличением количества камер видеонаблюдения, ручное наблюдение операторами стало менее эффективным. Автоматизированный анализ записей с камер наблюдения необходим для своевременного обнаружения аномальных действий, таких как насилие, ограбления и другие виды девиантного поведения. Традиционные методы интеллектуальной обработки видеозаписей часто опираются на сверточные нейронные сети для извлечения пространственных характеристик, однако такие сети могут испытывать трудности с улавливанием дальних временных зависимостей на видео. Помимо этого, вычислительная мощность, необходимая для быстрой работы сверточных сетей увеличивается экспоненциально с ростом размера таких сетей.

В данной статье предлагается новая архитектура, использующая иерархические сети типа Transformer для обнаружения аномалий на видеозаписях с камер наблюдения без использования сверточных сетей. Подход использует возможности архитектуры Video Vision Transformer (ViViT) а также одного из подходов к самоконтролируемому обучению (self-supervised learning), называемому контрастным обучением для извлечения надежных векторных представлений из видеоданных, которые затем могут быть использованы в классических моделях для решения различных задач, таких как классификация, кластеризация и распознавание аномалий. Благодаря внедрению иерархической структуры была модель эффективно работает с видеозаписями переменной длины, позволяя улавливать как локальные, так и глобальные временные закономерности.

Разработанная архитектура иерархической нейронной сети типа Transformer обрабатывает видеофрагменты и объединяет их в векторные представления на уровне событий. Для обучения дискриминативным векторным представлениям используется контрастное обучение на уровне сегментов и событий без необходимости использования размеченных данных.

Эффективность данного подхода демонстрируется на реальных наборах данных, полученных с камер ви-

деонаблюдения, а результаты показывают значительное улучшение производительности и качества в задачах классификации, кластеризации и обнаружения аномалий.

Литературный обзор

Обнаружение аномалий на видеозаписях широко изучалось, причем методы варьировались от традиционных методов машинного обучения, таких как логистическая регрессия до моделей глубокого обучения, таких как трехмерные сверточные нейронные сети. В ранних подходах для выявления отклонений от нормального поведения использовались созданные вручную признаки и статистические модели [1]. С развитием технологий глубокого обучения сверточные нейронные сети стали доминирующим инструментом для выделения признаков при анализе видео [2]. Однако такие сети ограничены в улавливании долгосрочных временных зависимостей из-за присущего им фокуса на локальных пространственных шаблонах.

Нейронные сети типа Transformer исторически использовались для обработки естественного языка и получения представлений для текстовых данных [3], с целью решения таких задач как классификация, генерация и сегментирование текстов. Тем не менее, архитектура Transformer продемонстрировала большой потенциал и в задачах компьютерного зрения [4]. Модель Video Vision Transformer (ViViT) [5] расширяет модель Vision Transformer (ViT), предложенную для обработки статических изображений, на видеоданные, позволяя моделировать пространственно-временные характеристики без использования сверток. В последних работах трансформеры были использованы для распознавания действий и «понимания» видео, продемонстрировав свою способность улавливать глобальный контекст.

Контрастное обучение — это парадигма самоконтролируемого обучения, которая обучается возвращать векторные представления, различая похожие и непохожие пары объектов [6]. Такие методы, как SimCLR [7] и MoCo [8], достигли значительных результатов в получении векторных представлений для изображений. В анализе видео контрастное обучение позволяет получать надежные векторные представления, используя временную согласованность и аугментацию данных [9].

Материалы и методы

Сети Transformer используют механизмы самовнимания (англ. self-attention) для оценки значимости различных частей входных данных [3]. Это позволяет модели улавливать дальние зависимости, что очень важно для понимания временных закономерностей в данных, представленных в виде последовательностей.

Контрастное обучение направлено на получение векторных представлений путем противопоставления «положительных» и «отрицательных» образцов [6]. Добиваясь максимального согласия между различными измененными представлениями одной и той же точки данных, модель обучается инвариантным характеристикам, которые хорошо обобщаются для последующих задач.

Предлагаемый фреймворк состоит из двух основных компонентов:

1. компонент извлечения векторных представлений на уровне сегментов. На уровне сегментов используется архитектура ViViT с контрастным обучением для получения векторов из видеосегментов фиксированной длины;
2. компонент объединения сегментов на уровне событий. Иерархическая сеть Transformer объединяет сегментные векторы в комплексное векторное представление на уровне событий, тем самым обрабатывая видеозаписи переменной длины.

ViViT обрабатывает видеоданные, рассматривая их как последовательность участков изображений, извлеченных из кадров видео. Он моделирует пространственные и временные размеры с помощью механизмов внимания. Для данной модели применяется позиционное кодирование (англ. positional encoding) для сохранения временного порядка, а размерность результирующего векторного представления равна 1024.

Для контрастного обучения используется нормализованная функция потерь перекрестной энтропии с учетом температуры (NT-Xent Loss) [7]. Генерируя положительные пары с помощью изменения исходных данных (путем обрезки, изменения разрешения, зашумления, поворота изображения) и рассматривая другие образцы как отрицательные, модель обучается «приближать» похожие векторные представления друг к другу, и «отдалять» непохожие в признаковом пространстве.

Видеофрагменты переменной длины представляют собой проблему для моделей, оперирующих данными фиксированного размера. Эту проблема решается путем внедрения архитектуры иерархического трансформера.

Трансформер на уровне сегментов обрабатывает сегменты фиксированной длины (16 кадров) и возвращает векторные представления сегментов. Трансформер уровня событий агрегирует полученные представления сегментов для формирования векторов на уровне событий, фиксируя глобальные временные зависимости. Такой иерархический подход позволяет модели обрабатывать события различной длительности без потери важной информации.

Предварительное обучение модели ViViT на сегментном уровне проводилось на наборе данных DCSASS

[10], содержащем видеозаписи с камер видеонаблюдения с девиантным поведением. Обучение проводилось на четырех графических ускорителях NVIDIA A100 в течение 120 часов. После этого проведено обучение модели уровня событий, которое заняло около 200 часов.

Результаты

Использованные наборы данных:

1. набор данных DCSASS [10]: набор данных с камер видеонаблюдения с различными аномальными событиями, такими как ограбления, насильственные действия и девиантное поведение;
2. набор данных ShanghaiTech Campus [11]: содержит записи видеонаблюдения из кампуса шанхайского технического университета, размеченные на «нормальные» и «абнормальные».

Использованные метрики для оценки:

1. задача многоклассовой классификации: Precision, Recall и F1-score с макро-усреднением по классам;
2. задача кластеризации: коэффициент силуэта (silhouette);
3. задача обнаружения аномалий (сводится к задаче бинарной классификации): Precision, Recall и F1-score, с акцентом на метрику Recall.

Экспериментальная установка:

1. оборудование: четыре графических ускорителя NVIDIA A100 (40GB).
2. программное обеспечение: фреймворк PyTorch для программной реализации и обучения моделей.

Гиперпараметры:

1. размер получаемого векторного представления: 1024;
2. количество слоев в модели уровня сегментов: 12;
3. количество модулей внимания в модели уровня сегментов: 16;
4. количество слоев в модели уровня событий: 4;
5. количество слоев в модели уровня событий: 8;
6. размер партии обучения: 16.

Обсуждение

Использование метода кластеризации K-средних (K-Means) для векторных представлений на уровне событий показали метрику силуэта в районе 0.69 для двух кластеров. При постепенном увеличении количества кластеров до 15, метрика падает до значения 0.44. Ручная проверка подтвердила, что кластеры представляют собой значимые группы, которые можно описать как «нормальное поведение», «ограбления» и «аресты» и т.д. Высокие показатели метрики силуэта свидетельствуют о четко определенных кластерах, что подтверждает качество полученных векторных представлений.

Для решения задачи классификации необходима тонкая настройка модели (fine-tuning). После обучения классификатора для решения задачи многоклассовой классификации получены следующие метрики:

- Macro Precision: 0,71–0,76.
- Macro Recall: 0.79–0.84.
- Макро F1-score: 0,74–0,80.

ROC-кривые, построенные для каждого класса также продемонстрировали способность модели качественно отделять объекты одного класса от других, а высокие показатели метрики recall особенно важны для минимизации количества пропущенных аномальных событий.

Задача обнаружения аномалий в данном контексте сводится к задаче бинарной классификации видеорядов на «нормальные» и «аномальные». После процедуры тонкой настройки и обучения классификатора получены следующие метрики:

- Precision: 0.76–0.79.
- Recall: 0.86–0.89.
- F1-score: 0,81–0,85.

Акцент на метрике Recall соответствует цели снижения количества ложноотрицательных результатов, что гарантирует обнаружение большинства аномалий. Результаты показывают, что модель эффективно выявляет аномальные события среди большого потока нормальных.

Предложенная архитектура иерархического трансформера успешно улавливает как локальные, так и глобальные временные зависимости, не опираясь на использование сверточных нейронных сетей. Использование контрастного обучения повышает дискриминативную способность векторных представлений. Устойчивая производительность в различных задачах демонстрирует универсальность предложенного метода.

Заключение

В данной статье представлена иерархическая архитектура на основе нейронных сетей типа Transformer, которая служит для обнаружения аномалий в видеозаписях с камер наблюдения. Представленная система работает без использования сверточных нейронных сетей, что делает ее более вычислительно эффективной и производительной. Архитектура эффективно справляется с видеоклипами переменной длины и демонстрирует высокую производительность в задачах классификации, кластеризации и обнаружения аномалий.

Предлагаемая система может помочь операторам видеонаблюдения, автоматически обнаруживать аномальные действия, снижая когнитивную нагрузку и вероятность человеческой ошибки. Благодаря оповещению об обнаруженных аномалиях в реальном времени

сотрудники служб безопасности могут быстрее реагировать на инциденты, что позволяет предотвратить нанесение вреда или ущерба.

Автоматизированный анализ позволяет одновременно контролировать большее количество камер без привлечения дополнительных человеческих ресурсов, что повышает эффективность наблюдения и положительно сказывается на общественной безопасности.

В качестве дальнейшей работы можно выделить пути интеграции предложенной системы в существующие программные комплексы для видеонаблюдения, а также оптимизация предложенной системы для улучшения ее производительности и сокращения времени ее обучения.

Также стоит отметить необходимость тестирования работоспособности системы и на других данных, полученных из различных сред, в которых ведется видеонаблюдение.

ЛИТЕРАТУРА

1. Adam A. et al. Robust real-time unusual event detection using multiple fixed-location monitors //IEEE transactions on pattern analysis and machine intelligence. — 2008. — Т. 30. — №. 3. — С. 555–560.
2. Simonyan K., Zisserman A. Two-stream convolutional networks for action recognition in videos //Advances in neural information processing systems. — 2014. — Т. 27.
3. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
4. Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale //arXiv preprint arXiv:2010.11929. — 2020.
5. Arnab A. et al. Vivit: A video vision transformer //Proceedings of the IEEE/CVF international conference on computer vision. — 2021. — С. 6836–6846.
6. Hadsell R., Chopra S., LeCun Y. Dimensionality reduction by learning an invariant mapping //2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). — IEEE, 2006. — Т. 2. — С. 1735–1742.
7. Chen T. et al. A simple framework for contrastive learning of visual representations // International conference on machine learning. — PMLR, 2020. — С. 1597–1607.
8. He K. et al. Momentum contrast for unsupervised visual representation learning // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2020. — С. 9729–9738.
9. Qian R. et al. Spatiotemporal contrastive video representation learning // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2021. — С. 6964–6974.
10. DCSASS Dataset [Электронный ресурс] // Kaggle. URL: https://www.kaggle.com/datasets/mateohervas/dcsass-dataset (дата обращения 04.10.2024).
11. Luo W., Liu W., Gao S. A revisit of sparse coding based anomaly detection in stacked rnn framework // Proceedings of the IEEE international conference on computer vision. — 2017. — С. 341–349.

Журнал «Современная наука: актуальные проблемы теории и практики»