DOI: https://doi.org/10.15276/ict.02.2025.54
УДК 004.942

Архітектура стека технологій мультикласової детекції об'єктів у потоковому відео

Філін Дмитро Владиславович
Аспірант каф. Інженерії програмного забезпечення
ORCID: https://orcid.org/0009-0000-6228-3657; fdima0311@gmail.com
Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

АНОТАЦІЯ

Роботу присвячено вирішенню актуальної технічної задачі інтелектуальної обробки відео потоків високої роздільної здатності з мінімальною затримкою. Метою роботи є забезпечення високої точності детекції та класифікації об'єктів у відео потоку високої роздільної здатності. В основі запропонованого рішення лежить декомпозиція процесу обробки на незалежні модулі з використанням патерну "Producer-Consumer", що дозволяє ефективно розмежувати операції введення-виведення та інтенсивні обчислення. Запропоновано гібридний дворівневий підхід до побудови висновків, який поєднує швидкість та ефективність модуля детектції на основі бібліотеки YOLO8 та модуля спеціалізованої класифікаційної моделі для поглибленого аналізу на основі на основі бібліотеки TensorFlow. Для досягнення максимальної продуктивності та уніфікації середовища виконання моделей використовується ONNX Runtime. Клієнтський доступ реалізовано за допомогою високопродуктивного асинхронного веб-фреймворку FastAPI. Обґрунтовано, що розроблена структура представляє комплексну архітектуру технологічного стека, що спроможна ефективно вирішувати втановлені задачі.

Практична цінність роботи полягає створенні масштабованої та гнучкої архітектури на основі прогресивних наукових досягнень, сучасних алгоритмічних та програмних рішень, здатних ефективно вирішувати складні задачі інтелектуальної обробки відео потоків високої роздільної здатності з в умовах обмежених обчислювальних ресурсів та часу.

Ключові слова: детекція; локалізація; класифікація, глибоке навчання; машинне навчання; гібридні системи, архітектура інформаційних систем

Актуальність. У сучасному світі спостерігається стале зростання обсягів відеоданих, що надає поштовх розвитку системи аналітики в реальному часі. Ручний аналіз таких обсягів інформації є не тільки трудомістким, але й практично неможливим у режимі реального часу. Це створює гостру потребу в автоматизованих системах, здатних аналізувати відео потоки, ідентифікувати об'єкти та події, і надавати цінну інформацію для прийняття рішень. Сфери застосування таких систем надзвичайно широкі. Особливою увагою користуються такі системи в критично важливих напрямах: сферах безпеки, промислової автоматизації, транспорту та автономних систем [1–3]. Але широке розповсюдження таких систем стикається з низкою складних технічних та архітектурних проблем, які вимагають ретельного проектування та оптимізації: забезпечення продуктивності в реальному часі для відео потоків високої роздільної здатності, гнучкості, масштабованості та забезпечення заданої точності роботи системи.

Технічні труднощі обробки зображень високої роздільної здатності зумовлені фундаментальними властивостями архітектур глибокого навчання [4–6]:

– обчислювальна складність обробки зображень зростає щонайменше квадратично відносно просторових розмірів;

– обсяг пам'яті, необхідний для зберігання карт активацій під час прямого та зворотного поширення, також масштабується квадратично. Це швидко вичерпує обсяг відеопам'яті навіть найсучасніших GPU, роблячи неможливою обробку зображень високої роздільної здатності в одному пакеті.

Вузьке місце продуктивності є не лише апаратною, а й архітектурною проблемою. Фундаментальне припущення про необхідність щільної, рівномірної обробки всього простору зображення, притаманне багатьом моделям, є вкрай неефективним для розрідженої природи інформації у високоякісних зображеннях. Більшість пікселів зображення часто є фоном або надлишковою інформацією, тоді як об'єкти, що становлять інтерес, можуть займати лише незначну частину зображення. Архітектури, такі як стандартні CNN та DETR,

застосовують обчислювально дорогі операції (згортки, глобальну само-увагу) рівномірно до всіх пікселів або патчів. Це означає, що обчислювальна вартість масштабується з загальною кількістю пікселів, а не з кількістю релевантної інформації. Отже, корінь проблеми полягає в архітектурній невідповідності: застосування щільних обчислень до проблеми з розрідженою інформацією.

Подолання багатомасштабних викликів, визначених вище, сьогодні здійснюється, як правило, за допомогою складних архітектурних рішень [7, 8]. Так, ключовим проривом у детекції об'єктів різного масштабу стала концепція мережі пірамід ознак (Feature Pyramid Network, FPN). FPN – це компонент моделей глибокого навчання, зокрема архітектур виявлення об'єктів, призначений для поліпшення виявлення об'єктів різного масштабу. Це стало вирішальним кроком, що дозволив детекторам ефективно обробляти багатомасштабні об'єкти. Однак ключовим обмеженням FPN є односпрямований потік інформації.

Для подолання цього обмеження запропоновано мережу агрегації шляхів (Path Aggregation Network, PANet), яка розширила FPN, додавши другий, зворотній потік інформації. Це створює двоспрямований потік даних, дозволяючи ознакам низького рівня ефективно враховуватися сумісно з ознаками високого рівня, що є критично важливим для точної локалізації. Але цей ефект досягається завдяки збільшенню кількості параметрів та обчислювальної складності.

Ще однією модифікацією стала двоспрямована FPN (BiFPN). Ця архітектура вдосконалює PANet, вводячи ваги, що навчаються, для кожної вхідної карти ознак. Це дозволяє мережі виконувати зважене злиття ознак і вивчати важливість різних ознак на різних масштабах, досягаючи кращої продуктивності з меншою кількістю параметрів.

Хоча здається логічним, що одна модель має вирішувати обидві задачі (детекції та класифікації), як це реалізовано в розглянутих архітектурах, архітектурний підхід, що розділяє детекцію та класифікацію на два етапи, залишається надзвичайно актуальним і може забезпечити краще рішення для складних інформаційних систем. Це пов'язано з кількома фундаментальними причинами, які стосуються компромісів між точністю, швидкістю, гнучкістю та складністю системи.

Метою дослідження є забезпечення високої точності детекції та класифікації об'єктів великої варіативності у відео потоку високої роздільної здатності з дотриманням обмежень продуктивності шляхом розробки архітектури стека технологій мультикласової детекції об'єктів.

Постановка задачі. Нехай неперервний відео потік V являє собою послідовність зображень (кадрів) I_t, де $t \in \{1, 2, \ldots, T\}$ – індекс кадру в послідовності (або час). Кожен кадр I_t — це растрове зображення розміром H x W x R, де H – висота, W – ширина, R – кількість каналів кольору.

Для кожного кадру I_t з відео потоку V система повинна надати набір детекцій $D_t = \{d_{t,1}, d_{t,2}, \ldots, d_{t,Kt}\}$, де K_t – кількість виявлених об'єктів у кадрі t.

Кожна детекція $d_{t,k}$ для k-го об'єкта в кадрі t містить:

– обмежувальний прямокутник**:** $b_{t,k} = (x_{min}, y_{min}, x_{max}, y_{max})$, що визначає просторове розташування об'єкта на кадрі.

– клас об'єкта: $c_{t,k} \in C$, де $C = \{c_1, c_2, \ldots, c_M\}$ – множина всіх можливих класів об'єктів, які система здатна класифікувати.

– ймовірність: $\$_{t,k} \in [0, 1]$, що вказує на впевненість моделі в правильності детекції та класифікації цього об'єкта.

Для оцінки ефективності системи використовуються функції втрат L використовуэться вираз, що вимірює розбіжність між передбаченими вихідними даними D_t та істинними даними D^{gt}_t для кожного кадру. Зазвичай L є комбінацією декількох компонент.

В якості функції втрат розглядаються метрики, що базуються на порівнянні передбачених детекцій з істинними:

– Precision (Точність): Частка правильно виявлених об'єктів серед усіх виявлених.

– Recall (Повнота): Частка правильно виявлених об'єктів серед усіх істинних об'єктів.

– Mean Average Precision (mAP): Усереднене значення Average Precision по всіх класах, є однією з найпоширеніших метрик для детекції об'єктів.

– Frames Per Second (FPS): Швидкість обробки кадрів, критична для роботи в реальному часі.

За таких умов задача мультикласової детекції об'єктів у потоковому відео полягає у побудові моделі F: $V \rightarrow \{D_1, D_2, D_T\}$, яка для кожного кадру I_t відео потоку V повертає набір детекцій $D_t = b_{t,k}, c_{t,k}, s_{t,k}\}$, здатну максимально точно виявляти, класифікувати та, при необхідності, відстежувати об'єкти з заданого набору класів C в умовах динамічної зміни сцен, освітлення, масштабу та ракурсу. Модель повинна мати високу ефективність (високі Precision, Recall та mAP) та забезпечувати прийнятну швидкість роботи FPS для застосування в реальному часі.

Розробка стека технологій мультикласової детекції об'єктів у потоковому відео в реальному часі пов'язана з розробкою гнучкої архітектури, здатної підтримувати складні, багатоступеневі аналітичні задачі. Центральною темою при цьому є неминучий компроміс між точністю моделі та швидкістю її роботи: більш складні та точні моделі, як правило, вимагають більше обчислювальних ресурсів і часу на обробку одного кадру.

Відомо, що спроба побудувати архітектуру, засновану на єдиній моделі значно ускладнює та здорожує процес навчання [7, 9, 10]. Для вирішення цієї проблеми перспективним виглядає використання двоступеневого підходу, де перша модель (детектор) вирішує лише одне завдання – знайти всі об'єкти загального класу. Потім детектовані об’єкти передаються другій, спеціалізованій моделі (класифікатору), яка навчена розрізняти специфічні класи.

Крім цього, двостадійні моделі демонструють вищу точність, але є повільнішими за одностадійні. Цей принцип можна застосувати на архітектурному рівні. Такий підхід дозволяє досягти оптимального балансу: система залишається швидкою, оскільки основна частина роботи виконується легкою моделлю, і водночас точною, оскільки для фінальної класифікації залучається потужна спеціалізована модель.

При реалізації інформаційної системи розділення завдань є класичним інженерним підходом, який значно спрощує розробку, підтримку та оновлення системи, завдяки забезпеченню модульності та гнучкісті в MLOps (Machine Learning Operations). Цей підхід відповідає принципам мікросервісної архітектури, де кожен компонент системи є незалежним і може розвиватися окремо [11, 12].

Принцип модульності також є важливим якщо потрібно внести будь які зміни до процесу машинного навчання. В цьому випадку не потрібно перенавчати складну єдину модель, достатньо донавчити або замінити відносно невелику модель класифікації, що значно дешевше, швидше та безпечніше.

Таким чином, хоча одномодельні детектори є чудовим рішенням для багатьох завдань, що вимагають високої швидкості та детекції загальних класів, двоступеневий підхід із розділенням детекції та класифікації є більш потужним, гнучким та масштабованим рішенням для побудови складних, високоточних та легких у підтримці промислових систем відео аналітики.

Звертаючи увагу на викладене, для вирішення задачі мультикласової детекції об'єктів у потоковому відео пропонується багатостадійна архітектура у вигляді конвеєра, побудована за принципом патерна проектування “producer-consumer”, розробленим для максимальної пропускної здатності та низької затримки. Стек технологій мультикласової детекції об'єктів складається з наступних етапів.

Eman 1. Захоплення кадрів (Producer): виділений потік для захоплення кадрів з джерела відео та розміщенням їх у спільну обмежену чергу. Це відокремлює введення-виведення відео від обробки, запобігаючи блокуванню та втраті кадрів.

Джерелом відео в цьому випадку буде IP-камера, веб-камера або відеофайл. Основна технологія для захоплення відео: OpenCV.

Eman 2. Просторово-часова обробка (Consumer): основний модуль обробки, який споживає кадри з черги. Виконується повний, високоточний конвеєр просторової детекції. Стратегія побудови висновків складається з двох моделей: грубий прохід на зменшеному зображенні ідентифікує області інтересу (детекція), а точний прохід застосовує потужну обробку (класифікацію) лише до високороздільних вирізок цих областей, що значно економить обчислювальні ресурси. Після того, як об'єкт виявлено та відстежено, вирізане зображення об'єкта передається до спеціалізованої, високопродуктивної моделі дрібнозернистої класифікації.

Модель 1. Швидкий генераліст (Fast Generalist). На цьому етапі використовується високооптимізована модель YOLOv8, яка виконує роль швидкого «генератора пропозицій». Її завдання – максимально швидко ідентифікувати на кадрі всі об'єкти, що становлять інтерес, та виконати їх грубу класифікацію.

Модель 2. Повільний спеціаліст (Slow Specialist). Для кожного об'єкта, знайденого на першому етапі, з оригінального кадру вирізається відповідна область (обмежувальна рамка). Цей невеликий фрагмент зображення передається на вхід другій, більш складній та точній класифікаційній моделі. Ця модель виконує поглиблену, вузькоспеціалізовану класифікацію і може бути реалізована на основі PyTorch або TensorFlow.

Eman 3. Потокова передача результатів (Пост-процесор/Producer): завершальний модуль, який форматує результати детекції та трекінгу у вказаний вихідний формат і передає їх потоком до клієнтських застосунків.

Таким чином, збудована система промислового рівня для цього завдання – це не єдина монолітна модель, а ретельно організована композиція спеціалізованих компонентів (захоплення, груба детекція, точна детекція, трекінг, класифікація, потокова передача), кожен з яких оптимізований для своєї конкретної ролі. Сама архітектура є такою ж важливою, як і моделі, що в ній містяться. Обробка відео в реальному часі – це потокова проблема, а не пакетна. Це негайно пропонує патерн “producer-consumer” для асинхронної обробки вводу-виводу та обчислень для запобігання втраті кадрів. Основне завдання детекції є занадто дорогим для виконання на кожному кадрі, що вимагає часової стратегії: розріджена, важка обробка на ключових кадрах та дешеве поширення на проміжних. Важка обробка на ключових кадрах все ще має бути ефективною, що вимагає просторової стратегії: підхід «від грубого до точного», щоб уникнути обробки всього кадру високої роздільної здатності найпотужнішою моделлю. Загальних класів детектора може бути недостатньо, що вимагає модульної конструкції, де вторинний, спеціалізований класифікатор може бути викликаний для дрібнозернистого аналізу відстежуваних об'єктів. Поєднання цих елементів показує, що рішенням є системна архітектура, конвеєр асинхронних, спеціалізованих модулів, а не просто одна «найкраща» модель детекції об'єктів.

Майбутні горизонти та передові напрямки досліджень. Поза межами поточних найсучасніших моделей, галузь детекції об'єктів у відео високої роздільної здатності розвивається у напрямках, що стосуються не лише архітектури, а й усього життєвого циклу моделі – від навчання до розгортання та взаємодії з користувачем.

Розгортання великих, високоточних моделей на пристроях з обмеженими ресурсами є ключовим викликом, що вимагає ефективних напрямків подальшого розвитку. В цьому випадку спільне проєктування апаратного та програмного забезпечення забезпечить тіснішу інтеграцію між архітектурами нейронних мереж та базовим апаратним забезпеченням (NPU, спеціалізовані мікросхеми) для подальшої оптимізації продуктивності на ват.

Проектування моделей з урахуванням специфічних апаратних прискорювачів (наприклад, GPU, NPU) є важливим для максимізації приросту продуктивності від технік, таких як арифметика низької точності.

Висновки. У роботі наведено формальну постановку задачі мультикласової детекції об'єктів у потоковому відео; обґрунтовано та запропоновано архітектуру для мультикласової детекції об'єктів у потоковому відео в реальному часі. Представлене рішення системно вирішує ключові виклики, пов'язані з продуктивністю, гнучкістю та масштабованістю. Основним внеском є інтеграція кількох передових архітектурних патернів та технологій в єдину, злагоджену систему.

Використання гібридного двоступеневого конвеєра висновування забезпечує гнучкість та масштабованість, дозволяючи поєднувати швидкість загального детектора з точністю спеціалізованих класифікаторів.

Патерн “”roducer-consumer" гарантує відмовостійкість та ефективне використання ресурсів шляхом декомпозиції процесів вводу-виводу та обчислень.

Представлена архітектура є міцною основою, яку в подальшому можна розширювати та вдосконалювати.

СПИСОК ЛІТЕРАТУРИ

1. Aharon N., Orfaig R., Bobrovsky B. “BoT-SORT: Robust Associations Multi-Pedestrian Tracking”. *arXiv*. 2022. DOI: https://doi.org/10.48550/arXiv.2206.14651.

2. Akyon F. C., Altinuc S. O., Temel L.” Slicing Aided Hyper Inference and Fine-tuning for Small Object Detection”. *2022 IEEE International Conference on Image Processing (ICIP)*. Bordeaux, France. 2022. p. 966–970. DOI: https://doi.org/10.1109/ICIP46576.2022.9897990.

3. Cao J., Weng X., Kitani K. “Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking”. *arXiv*. 2022. DOI: https://doi.org/10.48550/arXiv.2203.14360.

4. Lin T.-Y., Dollár P., Girshick R., He K. “Feature Pyramid Networks for Object Detection”. *arXiv*. 2016. DOI: https://doi.org/10.48550/arXiv.1612.03144.

5. Liu S., Qi L., Qin H., et al. “Path Aggregation Network for Instance Segmentation”. *arXiv*. 2018. DOI: https://doi.org/10.48550/arXiv.1803.01534.

6. Liu Z., Lin Y., Cao Y., et al. “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”. *arXiv*. 2021. DOI: https://doi.org/10.48550/arXiv.2103.14030.

7. Tan M., Pang R., Le Q. V. “EfficientDet: Scalable and Efficient Object Detection”. *arXiv*. 2019. DOI: https://doi.org/10.48550/arXiv.1911.09070.

8. Wang C.-Y., Yeh I.-H., Liao H.-Y. M. “YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information”. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2402.13616.

9. Wang C.-Y., A. B. L. et al. “YOLOv10: Real-Time End-to-End Object Detection”. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2405.14458.

10. Wojke N., Bewley A., Paulus D. “Simple Online and Realtime Tracking with a Deep Association Metric”. *arXiv*. 2017. DOI: https://doi.org/10.48550/arXiv.1703.07402.

11. Zhang Y., Sun P., Jiang Y, et al. “ByteTrack: Multi-Object Tracking by Associating Every Detection Box”. *arXiv*. 2021. DOI: https://doi.org/10.48550/arXiv.2110.06864.

12. Zhu X., Su W., Lu L., and others. “Deformable DETR: Deformable Transformers for End-to-End Object Detection”. *arXiv*.. 2020. DOI: https://doi.org/10.48550/arXiv.2010.04159.

DOI: https://doi.org/10.15276/ict.02.2025.54
UDC 004.4’24

Architecture of the multi-class object detection technology stack in streaming video

Dmytro V. Filin[1)]
Postgraduate student of the Department of Software Engineering
ORCID: https://orcid.org/0009-0000-6228-3657; fdima0311@gmail.com
[1)] Odesa Polytechnic National University, 1, Shevchenko Ave. Odesa, 65044, Ukraine

ABSTRACT

In modern blockchain systems, smart contracts are one of the most critical components for ensuring the automated execution of agreements without the need for intermediaries. However, smart contracts written in languages like Solidity may contain vulnerabilities that can be exploited by malicious actors to steal funds or manipulate assets. Given the increasing number of attacks on smart contracts, the development of effective methods for detecting such vulnerabilities is crucial. Traditional approaches to detecting vulnerabilities in smart contracts include symbolic execution, fuzzing, formal verification, and pattern matching. These methods have their advantages but face several challenges, such as high resource consumption, limitations in detecting new types of vulnerabilities, and difficulties in scaling to large contracts. As a result, there is a need to introduce new approaches, such as natural language processing (NLP) and machine learning, which can address these challenges more effectively. In this study, an NLP-based method was explored, using Word2Vec to convert smart contract code into vector representations, allowing for better analysis of the semantic relationships between elements of the code. These vector representations are then fed into a bidirectional recurrent neural network with GRU blocks and an attention mechanism. This approach allows the model to focus on the most important parts of the code and improve the accuracy of vulnerability detection. The comparative analysis showed that NLP-based methods significantly outperform traditional approaches in all key metrics. In particular, the GRU model with an attention mechanism demonstrated high results in accuracy, recall, and F-measure, making it effective for detecting complex vulnerabilities such as reentrancy. Furthermore, the NLP-based approach is capable of adapting to new types of attacks thanks to training on large datasets. Thus, the integration of NLP and machine learning represents a promising direction for enhancing the security of smart contracts. Future research can focus on improving these approaches, particularly through the implementation of advanced models such as transformers.

Keywords: Smart contract; Blockchain; deep learning; vulnerability detection; NLP; machine learning; symbolic execution; fuzzing; formal verification; pattern matching