m1p

A Preprint

Erik Ilyassov
Artificial Intelligence Center
Skolkovo Institute of Science and Technology
Moscow, Russia
E.Ilyassov@skoltech.ru

Svetlana Illarionova
Artificial Intelligence Center
Skolkovo Institute of Science and Technology
Moscow, Russia
S.Illarionova@skoltech.ru

Abstract

Задача детекции объектов в видеопотоках активно развивается благодаря достижениям глубокого обучения, однако перенос современных моделей в производственные и эксплуатационные процессы связан с рядом трудностей. Среди них – ограниченность специализированных наборов данных, высокая стоимость аннотирования и существенные доменные сдвиги, возникающие в реальных условиях съёмки. В данной работе проводится сравнение современных архитектур для видео-детекции, включая end-to-end трансформерные подходы и эффективные one-stage решения. Эксперименты выполняются как на открытых бенчмарках, так и на новом доменном наборе данных, содержащем объекты из промышленных сценариев. Такой дизайн позволяет оценить устойчивость и применимость актуальных методов компьютерного зрения в условиях реальных промышленных процессов.

1 Introduction

Автоматизированная видеоаналитика в производственных и эксплуатационных процессах является ключевым инструментом обеспечения безопасности, контроля качества и операционной эффективности. В отличие от статичных сцен, промышленные видеопотоки характеризуются высокой динамикой, множественными окклюзиями, бликами, пылью, сменой освещения и жёсткими требованиями к задержке. Система должна надёжно обнаруживать объекты, корректно определять их границы и поддерживать идентичность экземпляров во времени. При этом существенными ограничениями остаются дефицит специализированных данных и высокая цена аннотирования.

Современные методы видео-детекции объектов развиваются в двух направлениях. Первое связано с end-to-end трансформерными архитектурами: TransVOD и TransVOD++ агрегируют пространственно-временные признаки на уровне object-queries и устраняют зависимость от внешнего оптического потока. Второе направление — практичные one-stage решения, включая новые версии семейства YOLO, которые обеспечивают управляемый компромисс между точностью и скоростью в онлайн-режимах. В качестве бэкбонов используются ViT и Swin, а также их самосупервизорные варианты, позволяющие эффективно адаптировать модели к ограниченным доменным данным. Для оценки применимости подходов используются открытые наборы данных вроде ImageNet-VID и OVIS.

Несмотря на прогресс, перенос этих моделей на промышленные видеопотоки сопровождается рядом трудностей. Проблемой остаётся репрезентативность данных: существующие открытые наборы слабо покрывают промышленные классы и типичные визуальные артефакты, а их распределения часто дисбалансны. Даже сильные модели демонстрируют снижение устойчивости при длительных окклюзиях, схожести объектов по внешнему виду или глобальных доменных сдвигах. Дополнительно оффлайн-ориентированные архитектуры требуют значительных ресурсов и не всегда удовлетворяют ограничениям по латентности и стабильности FPS.

В данной работе формируется воспроизводимое сравнение современных моделей видео-детекции с упором на их промышленную пригодность. В качестве архитектурных линий рассматриваются TransVOD/TransVOD++ как представители end-to-end трансформеров, эффективные one-stage схемы (YOLO) с бэкбонами ViT/Swin и двумя конфигурациями SSL Swin. В качестве экспериментальной базы используются как открытые наборы данных (ImageNet-VID, UA-DETRAC, BDD100K video), так и новый датасет, включающий несколько реальных производственных сценариев. Его использование позволяет проверить применимость современных моделей видео-детекции в условиях, которые существенно отличаются от открытых бенчмарков по составу классов и визуальным особенностям. Для проверки выбранных моделей мы воспроизводим базовые результаты на открытых наборах данных, а затем адаптируем пайплайны к доменным условиям.

Итогом работы станут практические рекомендации по применению существующих подходов для задач интеллектуального мониторинга и контроля в промышленных сценариях. Это позволит выявить архитектурные решения, наиболее устойчивые к специфическим артефактам производственных видеопотоков, и зафиксировать роль качества и выбора данных как критического фактора итоговой устойчивости.

2 Headings: first level

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula. See Section 2.

2.1 Headings: second level

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

$$\xi_{ij}(t) = P(x_t = i, x_{t+1} = j|y, v, w; \theta) = \frac{\alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})} \tag{1}$$

2.1.1 Headings: third level

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Paragraph Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

3 Examples of citations, figures, tables, references

3.1 Citations

Citations use `natbib`. The documentation may be found at

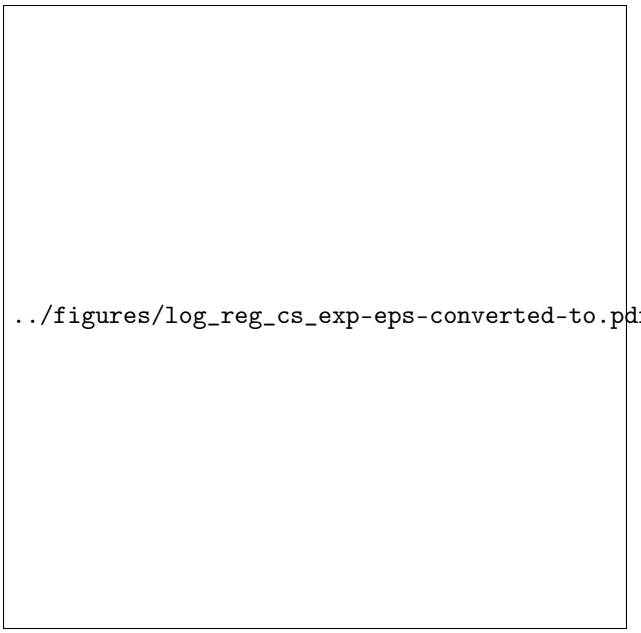

Рис. 1: Sample figure caption.

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Here is an example usage of the two main commands (`citet` and `citep`): Some people thought a thing [Kour and Saabne, 2014a, Hadash et al., 2018] but other people thought something else [Kour and Saabne, 2014b]. Many people have speculated that if we knew exactly why Kour and Saabne [2014b] thought this...

3.2 Figures

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi. See Figure 1. Here is how you add footnotes. [1] Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

3.3 Tables

See awesome Table 1.

The documentation for `booktabs` ('Publication quality tables in LaTeX') is available from:

`https://www.ctan.org/pkg/booktabs`

3.4 Lists

- Lorem ipsum dolor sit amet
- consectetur adipiscing elit.
- Aliquam dignissim blandit est, in dictum tortor gravida eget. In ac rutrum magna.

[1] Sample of the first footnote.

Таблица 1: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

Список литературы

George Kour and Raid Saabne. Real-time segmentation of on-line handwritten arabic script. In Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on, pages 417–422. IEEE, 2014a.

Guy Hadash, Einat Kermany, Boaz Carmeli, Ofer Lavi, George Kour, and Alon Jacovi. Estimate and replace: A novel approach to integrating deep neural networks with existing applications. arXiv preprint arXiv:1804.09028, 2018.

George Kour and Raid Saabne. Fast classification of handwritten on-line arabic characters. In Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of, pages 312–318. IEEE, 2014b. doi:10.1109/SOCPAR.2014.7008025.