

Отчёт о выполнении тестового задания
«DINO-Tracker: Taming DINO for Self-Supervised Point
Tracking in a Single Video»

Выполнил: Дюжев В. Д.

Санкт-Петербург, 2024

ОГЛАВЛЕНИЕ

Введение	1
1 Краткое описание	1
2 Постановка задачи	1
3 Структура отчета	1
Запуск предобученной модели	2
1 Базовая визуализация	2
2 Метрики качества	3
3 Производительность	5
4 Мульти-трекинг	6

Введение

Краткое описание

Dino-Tracker — метод трекинга, созданный на основе DINOv2-ViT, которая используется как базовая модель для получения качественных векторов-обобщений для изображений, содержащих важную семантическую информацию. Он позволяет добиться высокой точности, при этом относится к области self-supervised learning, что расширяет возможности его использования (т.к. нет необходимости разметки данных). Перед применением предполагается обучение на конкретном видео.

Постановка задачи

Дана последовательность кадров $\{I^t\}_{t=1}^T$, где T — длина видео. Задача состоит в обучении модели-трекера Π , принимающего на вход целевую точку x_q (query point) и предсказывающего набор $\{\hat{x}^t\}_{t=1}^T$ — траекторию данной точки (положения на всех кадрах). Целевая точка при этом задается положением и кадром (далее — начальный кадр). Таким образом итоговым результатом является возможность предсказания траектории любой точки любого кадра на протяжении всего видео.

Структура отчета

Данный отчет состоит из 5 секций:

1. **Запуск предобученной модели** — содержит базовые результаты экспериментов с демонстрацией основных возможностей метода на предложенных в оригинальном репозитории данных.
2. **Принцип работы** — описание архитектуры нейросетевой модели и алгоритмов постобработки, а также процесса обучения.
3. **Запуск на собственных данных** — результаты обучения модели на собранных вручную реальных данных.
4. **Техническая информация** — описание процесса взаимодействия с ПО, дополнения к оригинальному репозиторию.
5. **Выводы** — подведение итогов, выдвижение гипотез для улучшения метода.

Запуск предобученной модели

В данной секции представлены результаты использования модели DINO-Tracker на последовательностях из датасета Tapvid-Davis-480. Данные доступны по ссылке, указанной в репозитории.

Базовая визуализация

В качестве примеров для демонстрации работоспособности базовой модели DINO-Tracker (предсказаний нейросетевой части, без постобработки) были выбраны видео-последовательности с достаточно простым профилем движения, без явных окклузий. На рисунках 1-3 изображены по 3 кадра (первый, центральный и последний) с демонстрацией промежуточных траекторий для целевых точек (начальным кадром во всех случаях является первый).



Рис. 1. Траектории DINO-Tracker. Tapvid-Davis-480/29.



Рис. 2. Траектории DINO-Tracker. Tapvid-Davis-480/26.



Рис. 3. Траектории DINO-Tracker. Tapvid-Davis-480/21.

Можем заметить, что полученные траектории имеют естественный вид. Базовый метод может быть успешно применен и к трекингу в более сложных окружениях (рисунок 4), однако важно чтобы целевая точка находилась в достаточно уникальной части изображения и не подвергалась окклюзиям. Например, на рисунке 5 видно, что целевая точка сопоставляется неправильно с определенного момента времени (из-за появления схожего объекта и окклюзии вервоначальной цели). Данные ограничения помогает во многом снять постобработка с предсказанием окклюзий, о которой будет сказано в дальнейшем.



Рис. 4. Траектории DINO-Tracker. Tapvid-Davis-480/13.

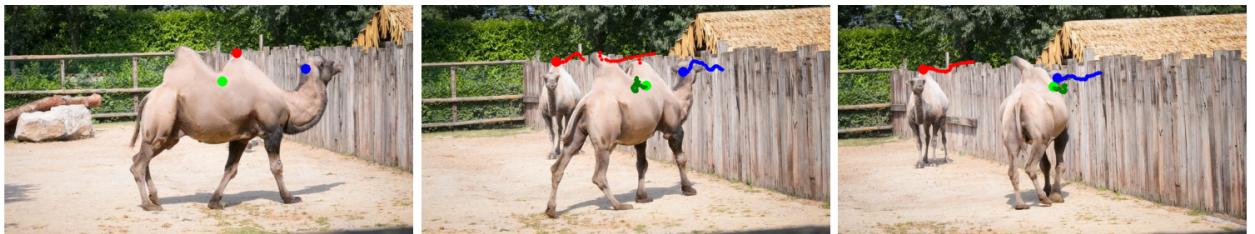


Рис. 5. Траектории DINO-Tracker. Tapvid-Davis-480/15.

Метрики качества

Оценка качества трекинга производится на датасете Tapvid-Davis-480 согласно инструкции, указанной в репозитории (также есть возможность провести оценку на других предложенных наборах данных, однако ввиду длительности процесса был выбран один из них).

Для оценки используются метрики, предложенные в оригинальной статье «TAP-Vid: A Benchmark for Tracking Any Point in a Video»:

1. occlusion accuracy (OA) — отношение количества правильно предсказанных окклюзий в последовательности к длине последовательности (стандартная accuracy классификации);
2. position accuracy ($< \delta^x$) — отношение количества правильно предсказанных положений видимых точек (без окклюзий) последовательности к количеству видимых точек, где правильными считаются пред-

сказания, лежащие в заданной окрестности (в пикселях) действительных значений;

3. jaccard — объединенная метрика для оценки качества предсказаний окклюзий и положений, рассчитываемая как отношение количества правильно предсказанных положений видимых точек с соответствующим правильным предсказанием окклюзий последовательности к количеству видимых точек + количеству неправильно классифицированных невидимых точек или точек с неправильными предсказаниями положений, предсказанных видимыми.

Важно отметить, что метрики $< \delta^x$ и jaccard рассчитываются для всех изображений последовательности (кроме начальных для соответствующих ключевых точек) с учетом приведения к размеру 256x256. В качестве итоговых метрик выступают OA , average jaccard (AJ) и average position accuracy ($< \delta_{avg}^x$), где последние две рассчитываются как средние значения jaccard и $< \delta^x$ для окрестностей в 1, 2, 4, 8 и 16 пикселей.

id	OA	AJ	$< \delta_{avg}^x$	id	OA	AJ	$< \delta_{avg}^x$
25	0.86407	0.52461	0.71564	23	0.88914	0.70477	0.8492
7	0.84293	0.62454	0.82839	14	0.95139	0.64647	0.75761
20	0.83599	0.5811	0.77156	13	0.89146	0.59985	0.72462
1	0.83536	0.55336	0.77867	29	0.96237	0.78638	0.87435
26	0.92519	0.72271	0.84825	17	0.89871	0.61162	0.75912
5	0.87887	0.60303	0.8446	4	0.87391	0.66864	0.88173
2	0.69653	0.37687	0.72344	11	0.79149	0.57035	0.7935
19	0.82501	0.47515	0.70126	16	1.0	0.94087	0.96645
6	0.85056	0.63005	0.80656	10	0.94926	0.76182	0.86853
22	0.79057	0.54181	0.76219	3	0.91635	0.75579	0.8585
8	0.88787	0.69386	0.84002	21	0.98371	0.84777	0.90917
28	0.93551	0.73576	0.84587	24	0.94805	0.8512	0.9379
0	0.99963	0.64083	0.72042	15	0.96577	0.80356	0.87399
9	0.69041	0.29614	0.5251	18	0.81178	0.52368	0.71774
27	0.87957	0.62347	0.7801	12	0.99365	0.88775	0.93579
<i>avg</i>	<i>0.8855</i>	<i>0.65279</i>	<i>0.80668</i>	—	—	—	—

Таблица 1. Метрики качества. Датасет Tapvid-Davis-480.

В статье завлены следующие данные для рассматриваемого датасета: $OA = 0.881$, $AJ = 0.646$, $< \delta_{avg}^x = 0.804$. Эксперимент подтверждает достижение таких значений.

Производительность

Эксперименты производятся на устройстве Lenovo Legion 7 (AMD Ryzen 9 5900HX, NVIDIA GeForce RTX 3080 120W 16Gb, 32 Gb RAM).

Для получения данных о производительности в инференсе метод был применен ко всем последовательностям из датасета. Эксперимент проводился только для нейросетевой части DINO-Tracker. Дальнейшее предсказание окклюзий сильно зависит от контекста изображения. На вход модели подавалась одна ключевая точка, соответствующая центру первого кадра. Принятый размер батча: 1. Ниже приведен график, иллюстрирующий зависимость времени обработки видео от количества кадров в нем.

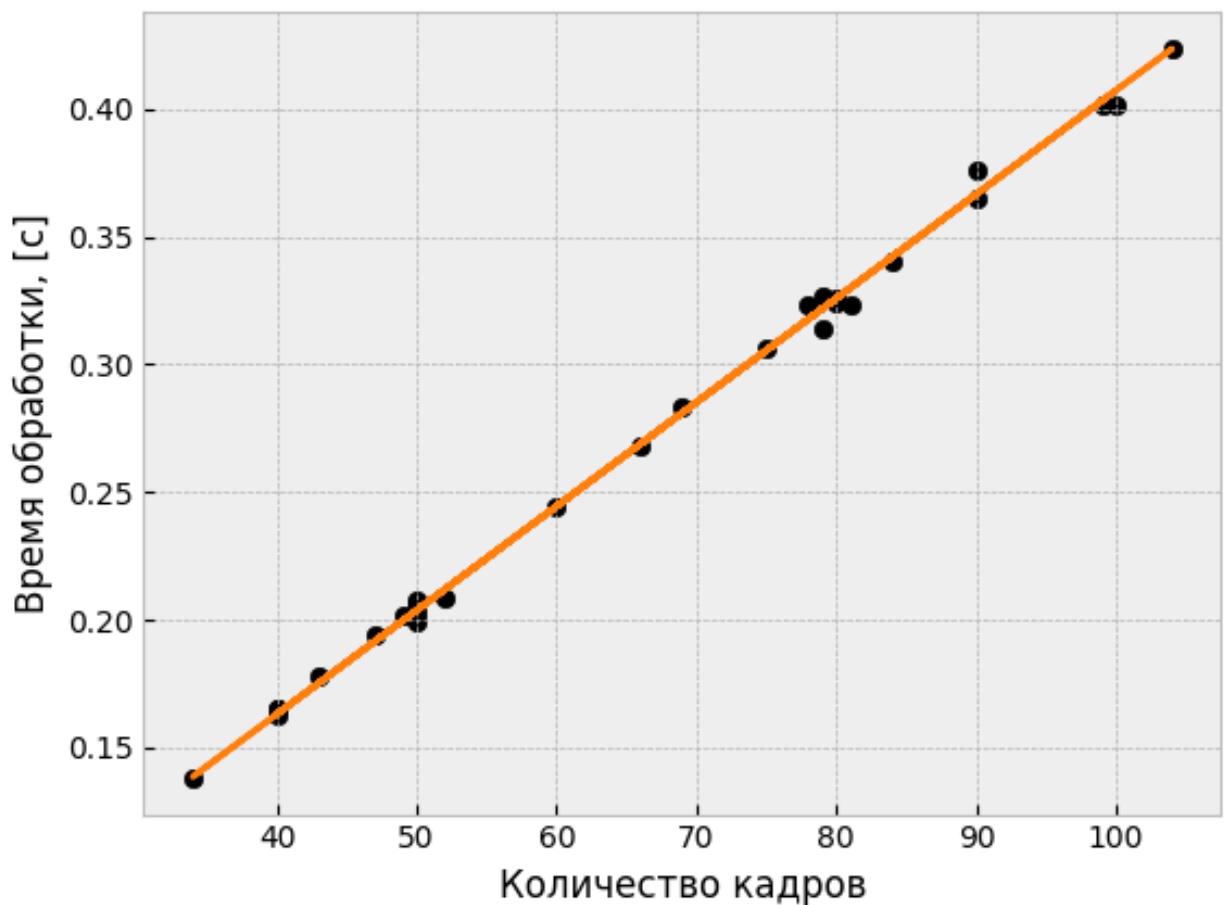


Рис. 6. Зависимость времени инференса нейросетевой части DINO-Tracker от количества кадров.

Алгоритм обработки видео последовательно предсказывает положение целевой точки на кадрах. Данные подтверждают, что зависимость является линейной. Аппроксимировав ее моделью линейной регрессии без смещения получим, что на предсказание позиции целевой точки в одном кадре тратится в среднем 0.0041 [с].

В процессе инференса в среднем затрачивалось 14.5 Гб видеопамяти.

Мульти-трекинг