



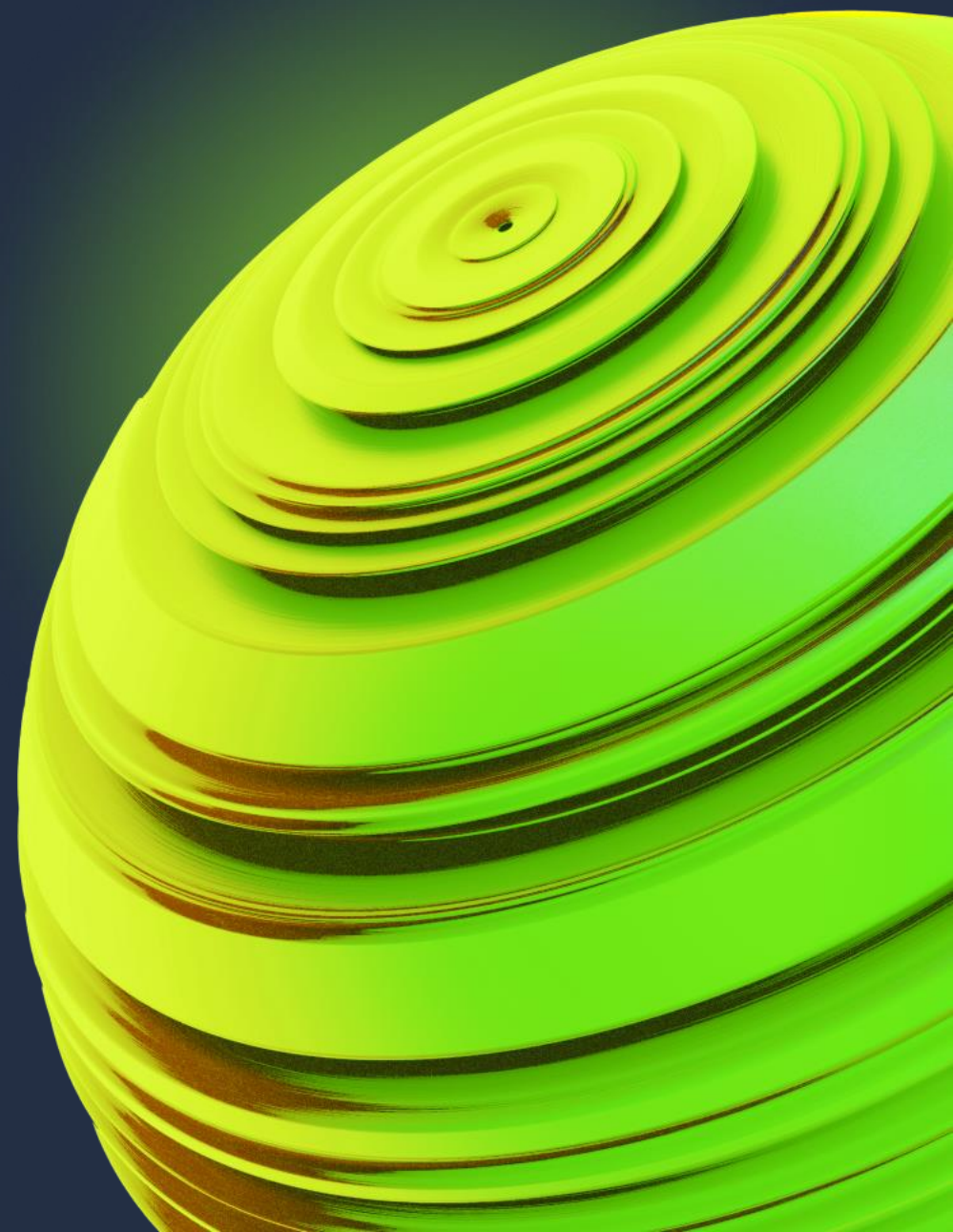
ИНСТИТУТ  
ДОПОЛНИТЕЛЬНОГО  
ОБРАЗОВАНИЯ  
УНИВЕРСИТЕТА ИННОПОЛИС



УНИВЕРСИТЕТ  
ИННОПОЛИС

## Распознавание действий на видео с использованием VideoSwin Transformer

Итоговое задание для курса “Профессия ML-инженер”





## Актуальность проекта

Актуальность систем распознавания действий на видео обусловлена растущим спросом в ключевых областях:

- видеонаблюдение и безопасность
- спортивная аналитика
- медиа аналитика
- робототехника

Современные архитектуры нейронных сетей позволяют эффективно учитывать пространственно-временной контекст, обеспечивая высокую точность и устойчивость к шуму.



## Цели и задачи проекта

Разработка модели для распознавания действий на видео, которая обеспечивает высокую точность предсказаний.

Основные задачи:

- выбор архитектуры
- подготовка и предобработка видеоданных
- обучение и оценка качества модели
- реализация веб-интерфейса



## Технологии и библиотеки

### Ядро разработки:

- [PyTorch](#) – реализация архитектур, обучение модели
- [TorchVision](#) – предобученные модели
- [OpenCV](#) – декодирование видео, работа с кадрами

### Оптимизация:

- [ONNX](#) – экспорт моделей, унификация инференса
- [ONNX Runtime](#) – получение предсказаний модели

### Интерфейс:

- [Gradio](#) – веб-интерфейс для инференса модели

### Вспомогательные инструменты:

- [Albumentations](#) – аугментация данных
- [NumPy](#) – препроцессинг данных



## Модели для распознавания действий

- **VideoSwin Transformer**

Использование self-attention механизма для обработки видео в виде последовательности пространственно-временных окон

- **C3D (Convolutional 3D Network)**

Расширение 2D-свёрток до 3D, чтобы обрабатывать временные зависимости между последовательными кадрами

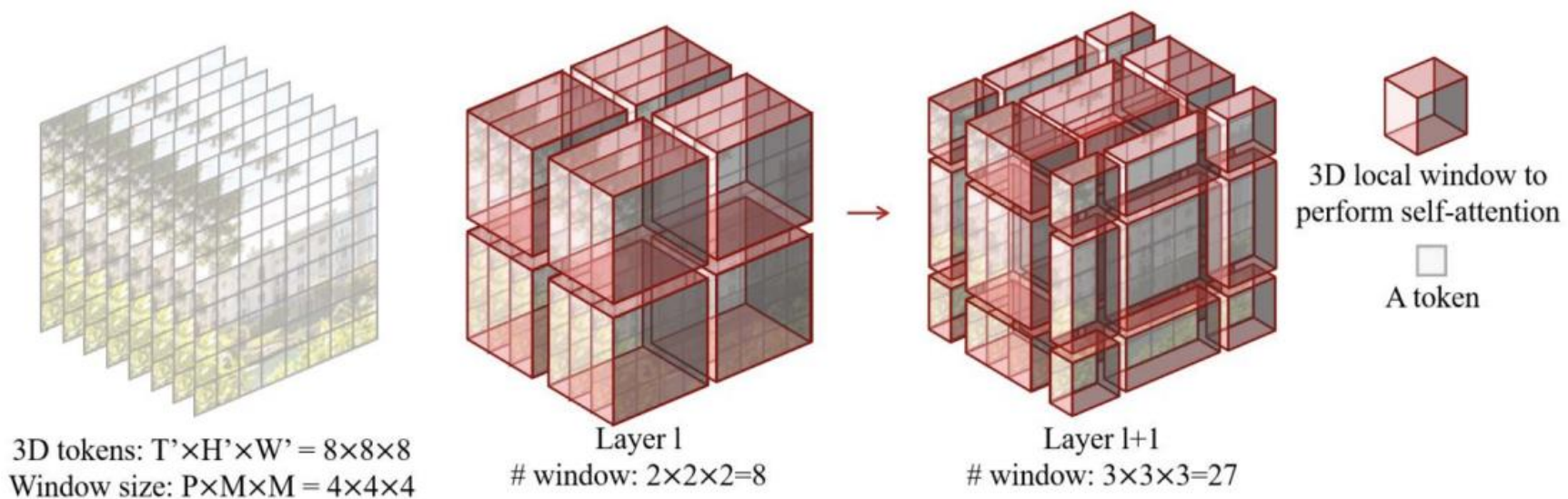
- **I3D (Inflated 3D ConvNet)**

Расширение предобученных 2D-сетей (например, Inception) до 3D с сохранением весов

- **R(2+1)D (Factorized Spatiotemporal Convolutional Networks)**

Разделение 3D-свёртки на последовательность 2D (пространственной) и 1D (временной) свёрток

# Архитектура VideoSwin Transformer



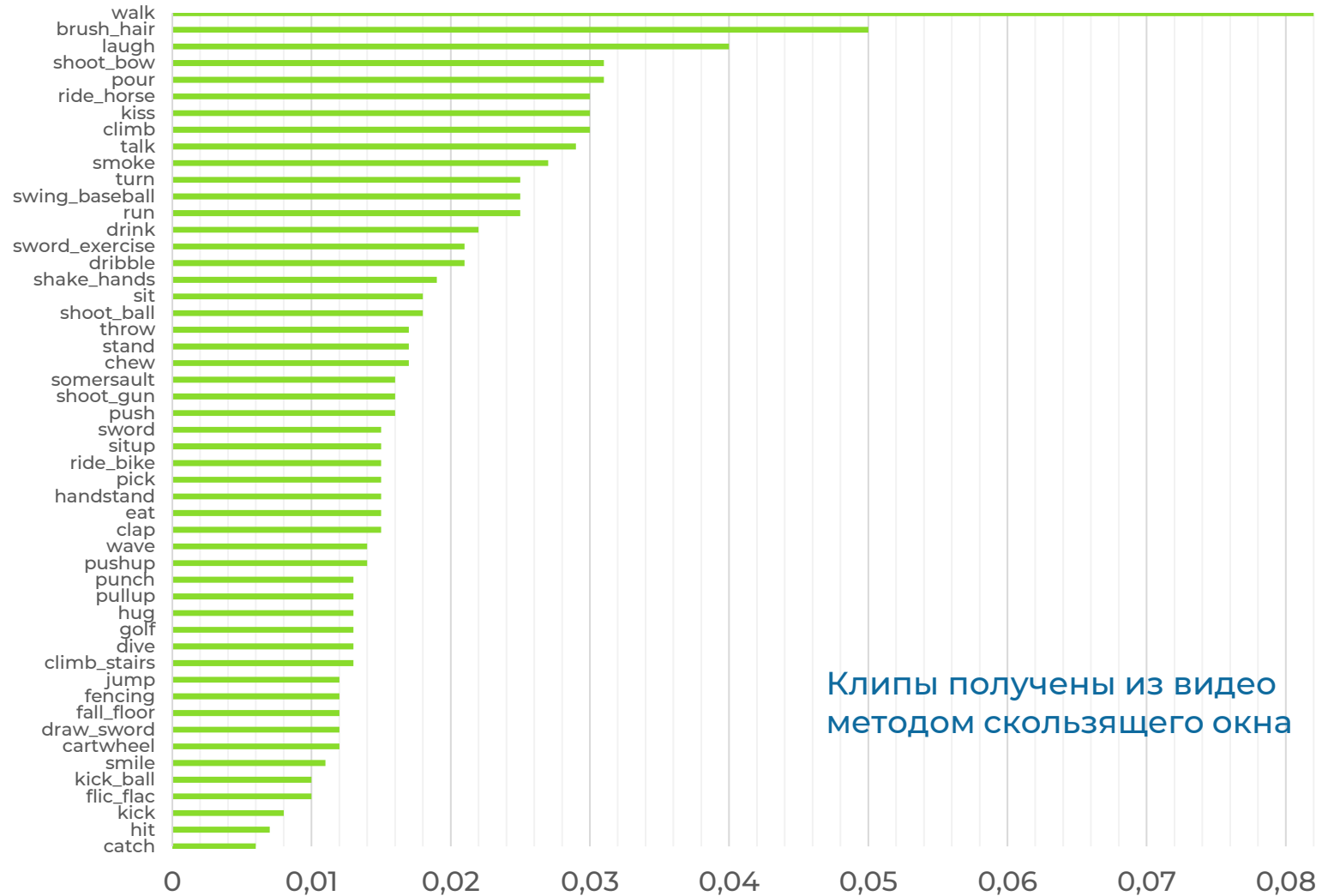
## Датасет HMDB51



- Содержит 51 класс действий, не менее 100 видеороликов на класс
- Разнообразие условий съёмки (ракурс, освещение, фон)
- Широко используемый бенчмарк для сравнения результатов



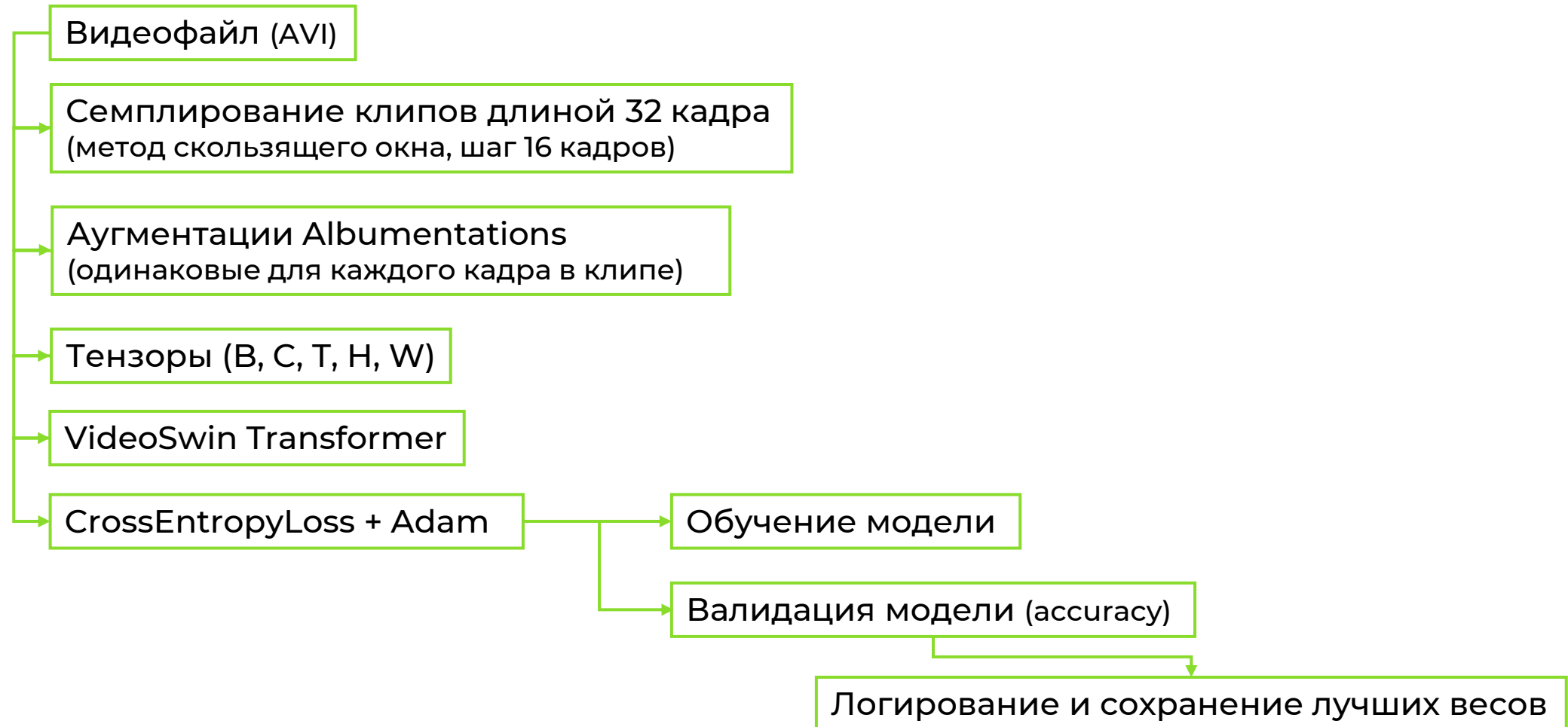
## Распределение клипов по классам





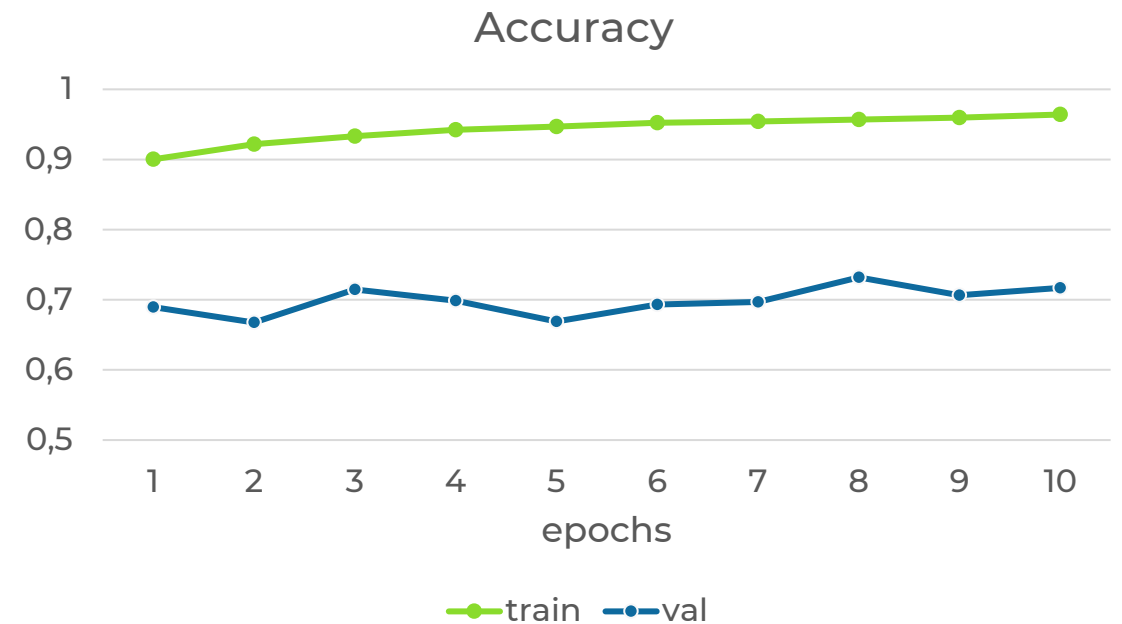
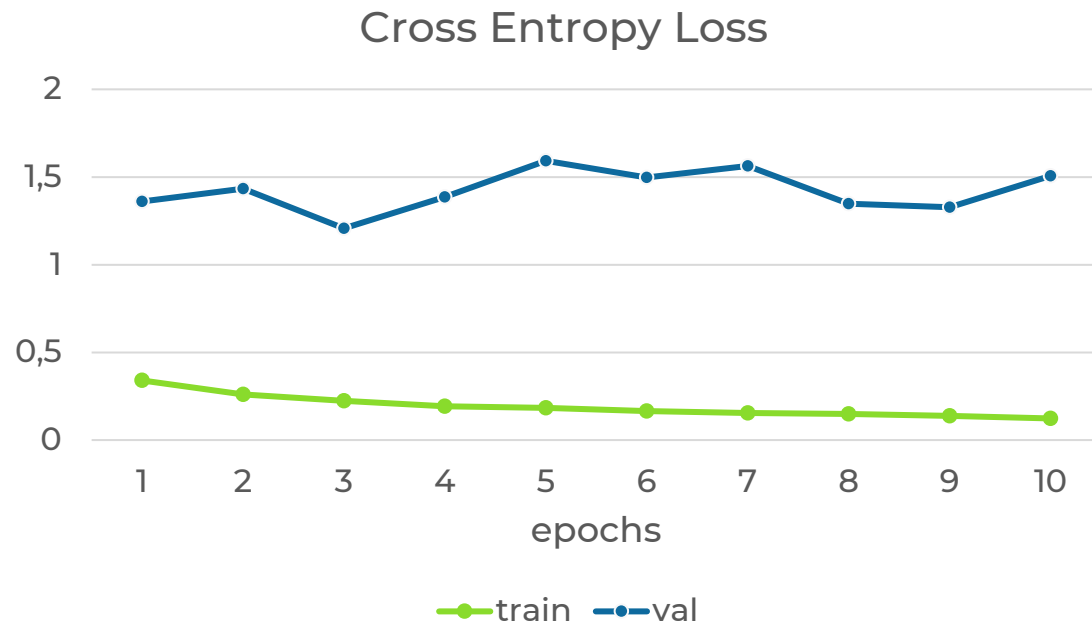


## Пайплайн обучения модели





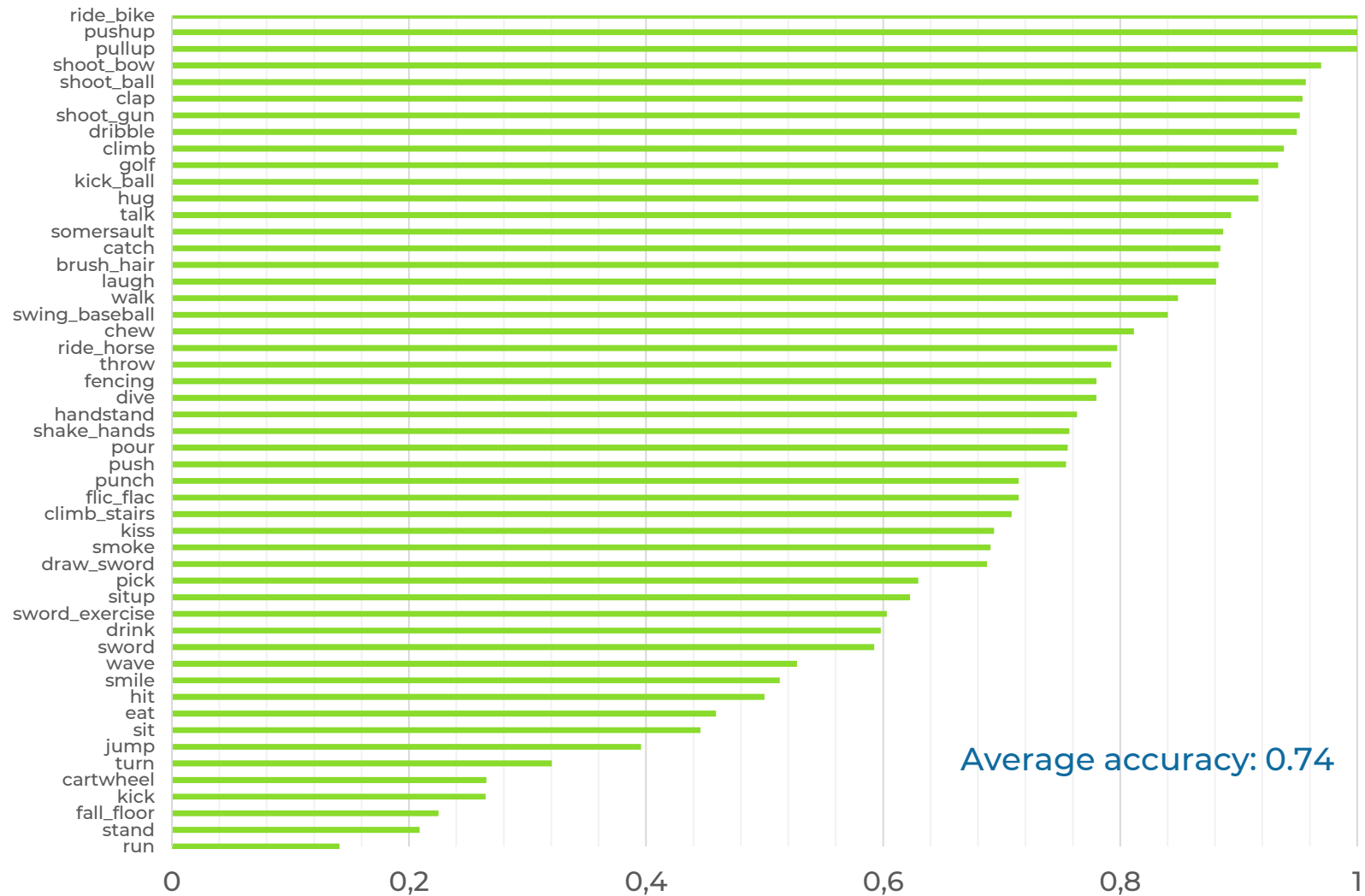
## История обучения модели



Обучение модели занимало 120 часов с использованием видеокарты NVIDIA RTX 4060Ti (16 GB)

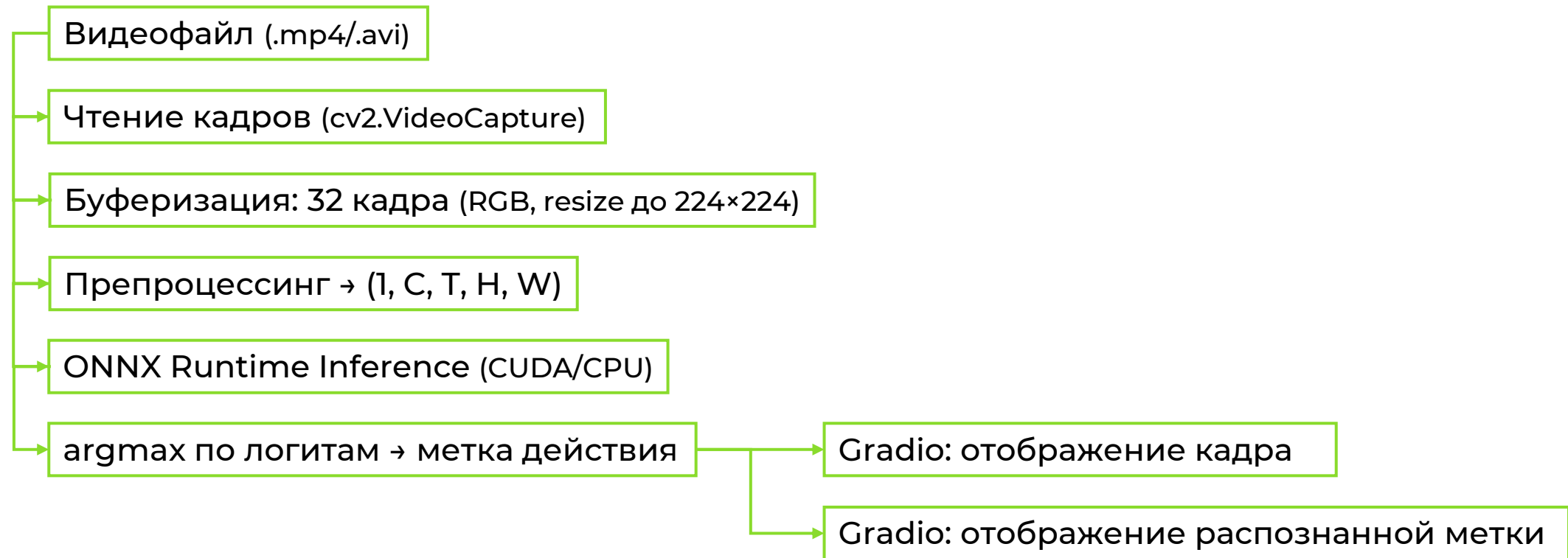


## Метрика на тестовой выборке (accuracy)





## Пайплайн инференса модели

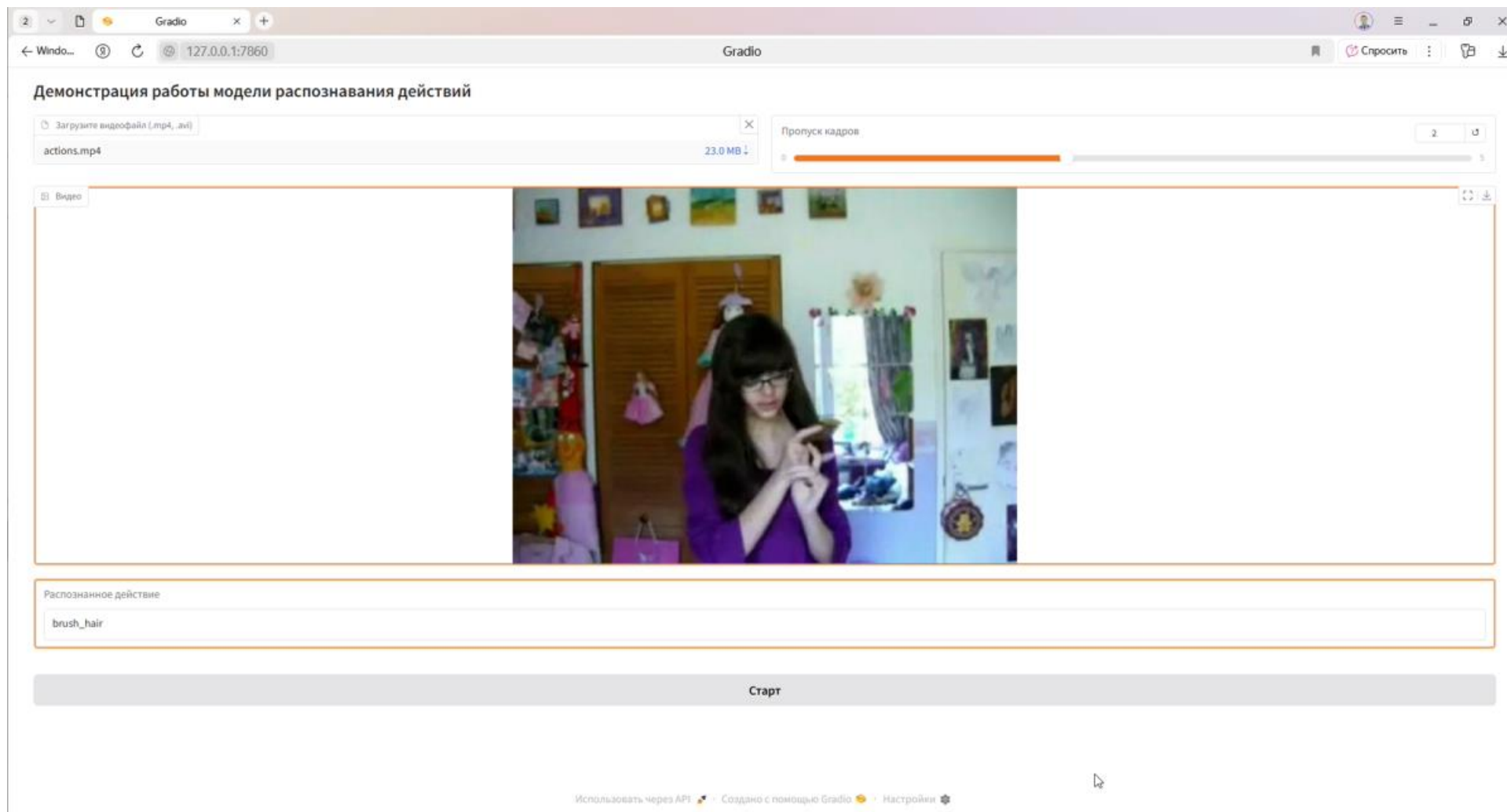




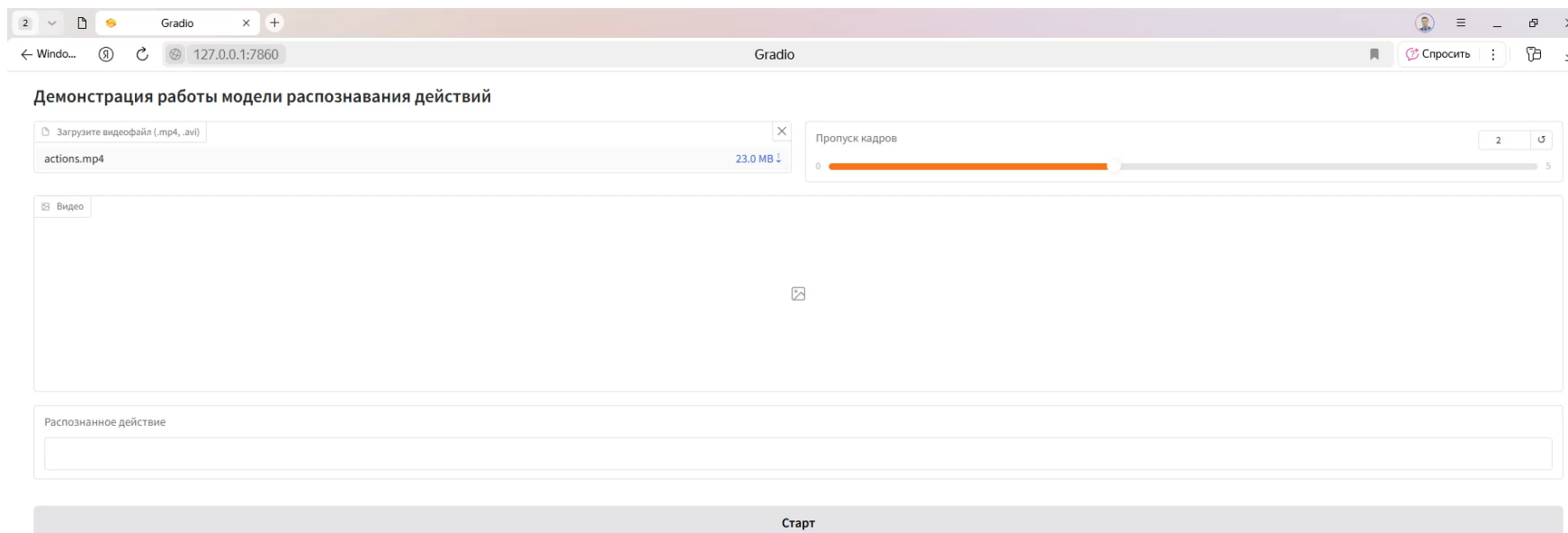
## Структура проекта

- [video\\_module](#) – Модуль со скриптами для работы пайплайна
- [split\\_data.py](#) – Разделение датасета на выборки
- [calc\\_label\\_distribution.py](#) – Анализ распределения видеоклипов по классам
- [train\\_model.py](#) – Обучение модели на основе VideoSwin Transformer
- [test\\_model.py](#) – Оценка точности модели на тестовой выборке
- [convert\\_model.py](#) – Конвертация модели из .pth в .onnx
- [run\\_model.py](#) – Запуск инференса модели на видеоролике
- [run\\_gradio.py](#) – Веб-интерфейс инференса с использованием Gradio
- [notebook.ipynb](#) – Jupyter-ноутбук для проверки работы пайплайна

# Интерфейс инференса модели



# Пример инференса модели (видео)



Использовать через API · Создано с помощью Gradio · Настройки



## Выводы

- Достигнута средняя accuracy ~74% на тестовой выборке датасета HMDB51
- Предобучение на видеодатасете Kinetic-400 обеспечило быструю сходимость и устойчивость к шуму и вариативности кадров
- Показано, что VideoSwin Transformer требует значительное количество ресурсов для обучения и инференса
- Подготовлен репозиторий проекта  
<https://github.com/i-a-elkin/MLInnopolis/tree/main/FinalAssessment>





ИНСТИТУТ  
ДОПОЛНИТЕЛЬНОГО  
ОБРАЗОВАНИЯ  
УНИВЕРСИТЕТА ИННОПОЛИС

Спасибо за внимание!


## Контакты

 +7(909) 052 97 36

 elkin@datalore.ru



## Сайт

 <https://github.com/i-a-elkin>

