**РАСПОЗНАВАНИЕ ДЕЙСТВИЙ НА ВИДЕО С ИСПОЛЬЗОВАНИЕМ VIDEOSWIN TRANSFORMER**

**1. ОБОСНОВАНИЕ ВЫБОРА ДАТАСЕТА И МЕТРИК**

**Выбор датасета HMDB51**

HMDB51 представляет собой один из наиболее репрезентативных и широко используемых датасетов для задачи распознавания действий на видео. Он включает 51 класс человеческих действий (например, прыжки, бег, кувырки), каждый из которых содержит не менее 100 видеороликов, снятых в различных условиях: съемка с разных ракурсов, при разном освещении, в разнообразных фонах и с разными качествами видео. Такая вариативность обеспечивает достаточный уровень сложности и баланс между реалистичностью данных и управляемым размером датасета. Кроме того, общеизвестные бенчмарки на основе HMDB51 позволяют сравнивать результаты с существующими работами, что важно для оценки конкурентоспособности предлагаемого решения.

**Выбор метрик**

В качестве основной метрики выбрана *accuracy* вычисленная для каждого класса (class-wise accuracy) и усредненная по всем классам (mean class accuracy). Такой подход имеет следующие преимущества:

* **Интерпретируемость результатов**: классическая точность (accuracy) интуитивно понятна и дает прямую оценку доли правильно классифицированных клипов.
* **Сопоставимость**: большинство работ по распознаванию действий используют mean class accuracy для HMDB51, что упрощает прямое сравнение.

**2. ПРЕДЛОЖЕННОЕ РЕШЕНИЕ**

В работе использовалась модель VideoSwin Transformer – прогрессивная архитектура на основе визуальных трансформеров, оптимизированная для видеоанализа. Ключевые особенности решения:

1. **Построчная многомасштабная обработка**. VideoSwin делит видео на окна (**windowed self-attention**) и динамически меняет масштаб пространственно-временного внимания, что повышает эффективность модели при анализе длинных видеофрагментов.
2. **Временная и пространственная иерархия**. Благодаря иерархической структуре модель последовательно агрегирует признаки от локальных к глобальным временным рамкам, что позволяет эффективно улавливать как мелкие детали, так и общие паттерны движения.
3. **Предобучение на крупных видео-датасетах**. Использование предобученных весов (например, на **Kinetics-400**) ускоряет процесс сходимости и значительно улучшает обобщающую способность модели при работе с новыми данными.

**3. АРХИТЕКТУРА И ОЦЕНКА СЛОЖНОСТИ МОДЕЛИ**

**Структура VideoSwin Transformer**:

* **Входной блок**: разделение видео на кубические блоки (tubelet embedding).
* **Иерархические слои**:
  + Каждый слой содержит блоки Swin Transformer с механизмом локального **self-attention**.
  + Между блоками происходит сдвиг окна для обеспечения перекрытия и агрегации контекстной информации.
* **Пулинги и уменьшение размерности**: после каждого слоя происходит снижение пространственного разрешения в два раза.
* **Классификационный слой**: **Global Average Pooling** по пространственно-временному объёму и полносвязный слой на 51 выход.

**Оценка вычислительной сложности**:

* **Параметры**: около 100M параметров для VideoSwin-Base, из которых ~80% приходится на блоки внимания.
* **Память**: требования к VRAM — около 16GB при batch size 2 клипа (в клипе 32 кадра).

**4. ДЕТАЛИ ОБУЧЕНИЯ**

1. **Предобработка**:
   * Масштабирование кадров до 224×224.
   * Аугментации: обрезка со случайным изменением размера, горизонтальный флип, изменение яркости/контраста, случайное поворотное искажение. Одинаковый набор аугментаций применялся ко всем кадрам в клипе.
2. **Гиперпараметры**:
   * Оптимизатор: Adam.
   * Скорость обучения: 1e-4.
   * Batch size: 2 клипа (в клипе 32 кадра) на GPU.
   * Длительность обучения: 10 эпох.
3. **Валидация и тест**:
   * По завершении каждой эпохи вычисляется mean class accuracy на валидационной подвыборке HMDB51.
   * Лучшая модель сохраняется по наивысшей accuracy.
4. **Оборудование**:
   * Обучение проводилось на NVIDIA RTX 4060TI (16GB VRAM).

**5. ВЫВОДЫ**

В ходе экспериментов VideoSwin Transformer продемонстрировал высокую эффективность при распознавании действий в разнообразных условиях датасета HMDB51. Основные выводы:

1. Достигнута средняя точность (**accuracy**) на уровне ~74%, что обусловлено эффективным механизмом **self-attention**, позволяющим модели улавливать сложные пространственно-временные зависимости в видеоданных.
2. Предварительное обучение на датасете **Kinetic-400** обеспечило быструю сходимость и устойчивость к шуму и вариативности кадров.
3. Несмотря на высокие требования к вычислительным ресурсам (как при обучении, так и при инференсе), использование механизма локальных окон существенно снижает вычислительную сложность операций внимания, что делает возможным обработку длительных видеофрагментов.

Таким образом, архитектура **VideoSwin Transformer** представляет собой эффективное и перспективное решение для задачи распознавания действий на видео, сочетающее высокую точность классификации с возможностью масштабирования под различные задачи и условия применения.