ДВУХПОТОКОВЫЙ ТРАНСФОРМЕР ДЛЯ РАСПОЗНАВАНИЯ ДЕЙСТВИЙ

Авазжон Марахимов, Кабул Худайбергенов
Кафедра Обработки информации и систем управления, Ташкентский государственный технический университет, Ташкент, Узбекистан
kabul.kudaybergenov@gmail.com

Аннотация. В данной работе предлагается новая архитектура нейронной сети для задачи распознавания человеческих действий на основе скелетных данных – Двухпотоковый Скелетный Трансформер. В качестве решения предлагается архитектура, в которой один поток (пространственный) специализируется на извлечении структурных паттернов с использованием топологически-ориентированного механизма внимания, а второй поток (временной) фокусируется исключительно на кинематике и динамике движений, обрабатывая производные координат (скорости и ускорения). Предложенный механизм перекрестного слияния позволяет эффективно объединять эти модальности, достигая высокой точности на стандартных наборах данных.

Ключевые слова: *распознавание действий, скелетные данные, нейронные сети, Трансформер, механизм внимания, графовая топология.*

1. Введение

Распознавание действий человека по скелетным данным является одной из ключевых задач компьютерного зрения с широким спектром применений: от систем видеонаблюдения и человеко-машинного взаимодействия до анализа спортивных тренировок и здравоохранения [1,2].

В данной работе предлагается **Двухпотоковый трансформер**, который концептуально разделяет задачу на две составляющие:

1) Понимание статической структуры и анатомических связей в каждом кадре.

2) Анализ чисто временной динамики и кинематики суставов независимо от их абсолютного положения.

2. Предлагаемая методология

Предлагаемая архитектура состоит из двух независимых ветвей: Потока топологии скелета и Потока динамики движений. На выходе оба потока объединяются через модуль кросс-внимания.

Пусть входная скелетная последовательность представлена тензором $X \in \mathbb{R}^{T \times V \times C}$, где T - количество кадров, V - количество суставов, C - количество каналов. Для Потока топологии используются абсолютные координаты суставов:

$$P = X \in \mathbb{R}^{T \times V \times C}$$

Для Потока динамики вычисляются движения (первая производная) и ускорения (вторая производная):

$$M_t = X_{t+1} - X_t$$

$$A_t = M_{t+1} - M_t$$

Входным тензором для временного потока служит конкатенация скоростей и ускорений $D = [M, A] \in \mathbb{R}^{T \times V \times 2C}$.

Пусть H_S - выход топологического потока, а H_T - выход динамического потока. В модуле кросс-внимания пространственные признаки используются как запросы (Queries), а временные динамические признаки как ключи (Keys) и значения (Values):

$$Q_F = H_S W_Q^F, \quad K_F = H_T W_K^F, \quad V_F = H_T W_V^F$$

$$H_{Fusion} = \text{Softmax}\left(\frac{Q_F K_F^T}{\sqrt{d_k}}\right) V_F + H_S$$

Такой подход позволяет модели “искать” в динамическом потоке те временные паттерны, которые релевантны для конкретных структурных поз, обнаруженных в пространственном потоке. Итоговый вектор H_{Fusion} пропускается через слой глобального усреднения и полносвязный слой (MLP) для классификации действия.

3. Вычислительные эксперименты

Ниже представлены результаты сравнения предложенного Двухпотокового скелетного трансформера (2S-SktFormer) с известными базовыми моделями на наборах данных NTU RGB+D 60 и NTU RGB+D 120. Значения для базовых моделей приведены согласно опубликованным источникам, значения для предложенной модели получены в результате собственных экспериментов.

Таблица 1. Сравнение точности распознавания действий (%).

Модель	**NTU-60 X-Sub**	**NTU-60 X-View**	**NTU-120 X-Sub**	**NTU-120 X-Set**	**Параметры (M)**
ST-GCN	81.5	88.3	70.7	73.2	3.1
2s-AGCN	88.5	95.1	82.9	84.9	6.9
ST-TR	89.9	96.1	82.7	84.7	12.1
CTR-GCN	92.4	96.8	88.9	90.6	5.8
DSTA-Net	91.5	96.4	86.6	89.0	8.4
2S-SktFormer	**93.1**	**97.2**	**89.6**	**91.2**	**5.2**

4. Заключение

Ключевая новизна заключается в строгом концептуальном разделении обработки. Предложенный метод перекрестного слияния эффективно интегрирует топологическую статику с кинематической динамикой. Развитие данного подхода открывает новые перспективы в создании более интерпретируемых и точных систем анализа движений.

Список литератур

1. Yan S., Xiong Y., Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition // Proceedings of the 32nd AAAI Conference on Artificial Intelligence. – 2018. – P. 7422-7429.

2. Vaswani A., Shazeer N., Parmar N. Attention is all you need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998-6008.