ENHANCING HUMAN-OBJECT INTERACTION DETECTION VIA SEMANTIC FEATURES AND NON QUERY-KEY-VALUE MECHANISMS IN TRANSFORMER MODELS

Pham Tan Tai
University of Information Technology
HCMC, Vietnam

What ?

We conducted a comprehensive study on human-object interaction (HOI) detection, in which we have:

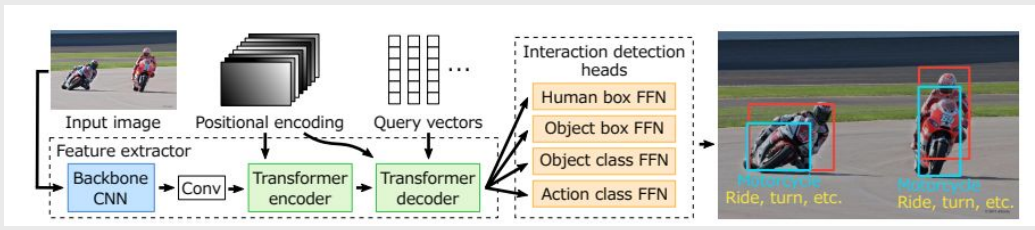
- Investigated state-of-the-art Transformer-based HOI models. Analyzed the use of semantic features to enhance interaction detection.
- Proposed an improved mechanism to integrate semantic features beyond the standard Query-Key-Value attention.
- Evaluated the proposed method on V-COCO and HICO-DET benchmarks.

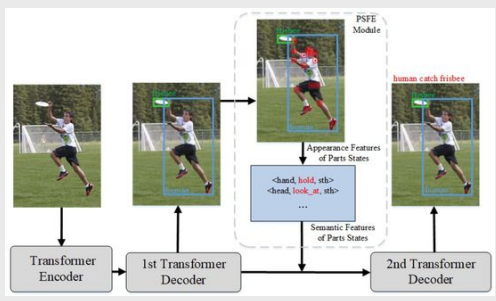
Why ?

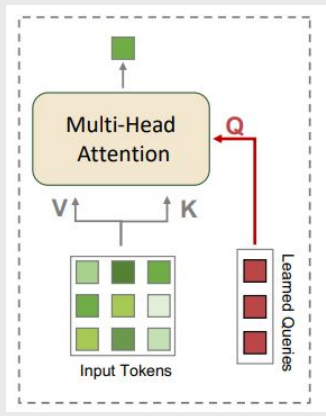
- Human-object interaction (HOI) is key to understanding human behavior in vision tasks and vital for applications like surveillance and autonomous driving.
- Current Transformer-based models still depend on standard QKV attention and often overlook semantic features in both encoder and decoder, limiting performance in complex real-world scenarios.

Overview

HOI Transformer pipeline | Semantic Injection between Decoders | Non QKV







Description

1. HOI Transformer pipeline

- The HOI Transformer pipeline begins with a convolutional backbone used to extract visual features from the input image.
- These features are then passed through positional encoding and fed into a Transformer encoder.
- Query vectors are initialized and passed into a Transformer decoder to interact with encoded visual information.
- Finally, interaction detection heads—each composed of feed-forward networks (FFNs)—predict the human box, object box, object class, and action class corresponding to each interaction.
- This allows the model to output structured triplets {human, object, interaction} for each detected instance.

2. Semantic Injection between Decoders

- To enhance contextual understanding, semantic features are integrated into Transformer decoder stages.
- In the decoder, these enriched features help disambiguate complex interactions by guiding query refinement and object-action reasoning.
- This dual-stage enrichment improves the model's ability to recognize subtle and diverse human-object interactions.

3. Non QKV

- Instead of traditional QKV interactions, learnable tokens replace either queries or keys to generate dynamic affinity matrices, focusing on query-only or key-only correlations.
- The resulting attention mechanism integrated into the Transformer architecture enhances feature representation for HOI tasks.
- This approach reduces computational complexity and improves adaptability to diverse visual inputs.

NII
Phạm Tấn Tài – Trường Đại học Công nghệ Thông tin. ĐHQG TP.HCM
TEL : 0858848302 Email : taipt.19@grad.uit.edu.vn