IMPROVING ACTION RECOGNITION IN HUMAN-OBJECT INTERACTION DETECTION PROBLEM USING MULTI-FEATURE FUSION FOR TRANSFORMER ARCHITECTURE

Lê Việt Thịnh [1,4] **Đinh Nhật Minh** [2,4] **Dương Thành Bảo Khanh** [3,4]

{[1] 20520781, [2]20521597, [3]20521444}@gm.uit.edu.vn [4]Trường Đại học Công nghệ Thông tin ĐHQG TP.HCM

What ?

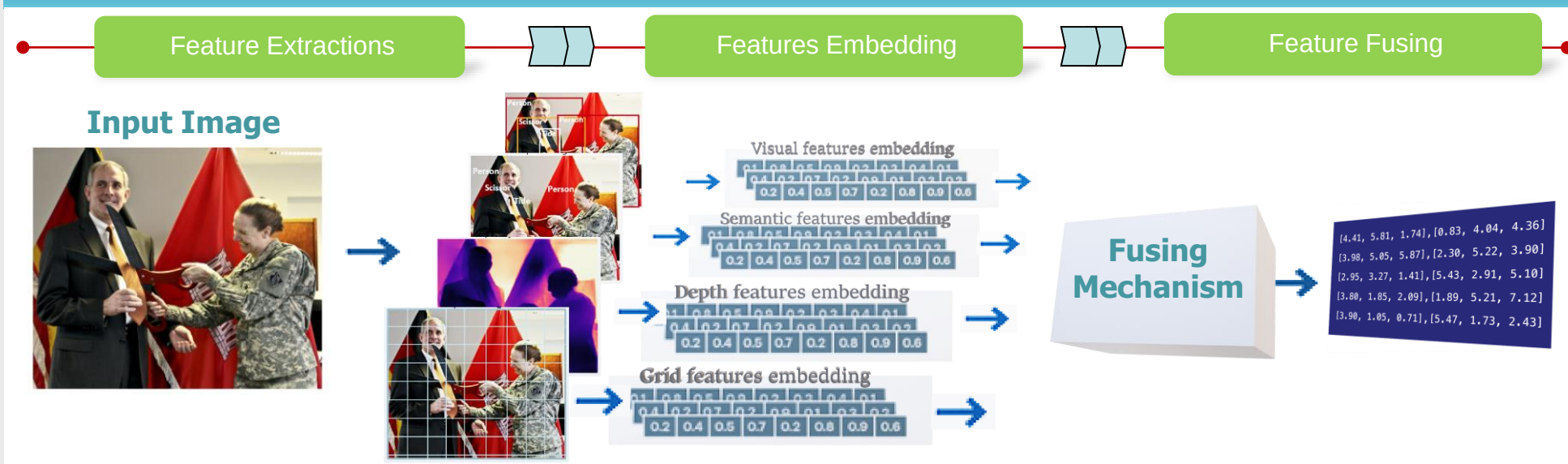
We introduce a framework to advance in human-object interaction detection problem, in which we have:

- Explored and analyzed the effectiveness of various features when applied to the problem of detecting human-object interactions.
- "Propose a novel method that integrates information from diverse features, which can be applied to a transformer architecture to achieve high performance on the V-COCO dataset."

Why ?

- The applications of human object detection are diverse and can be utilized in various fields of life. For instance, detecting anomalous behavior, supporting security and traffic monitoring. However, detecting human-object interactions is such a complex task, requiring a model that can recognize semantic information in images.
- Recent notable studies leveraging transformer have only focused on using visual features, while neglecting other valuable information like: spatial and semantics features

Overview

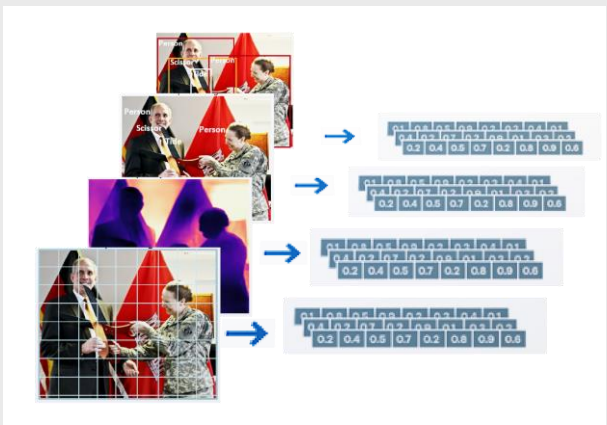


Description

1. Feature Extractions

- This Input image will pass through an object detector that recognise objects in the image. Additionally, features of the image will also be extracted.
- Sequentially from left to right and from top to bottom: Visual features, semantic feature, depth features, grid features extracted using grid-based approach.

- Here, we can use the embedding mechanism to embed all features that we are interested in, ensuring that the resulting embeddings retain the most important semantic information while being adaptable to different applications and use cases.



- This mechanism plays a crucial role in that it require capturing the essence of the input image.
- We planed to develope a reliable and efficient fusing mechanism that generate the final embedding vector capable of retaining the most relevant information from the feature embeddings, while being robust to noise and outliers in the data, therefore critical in enhancing the performance of many machine learning models

2. Feature embedding

- An embedding is a numeric representation derived from a extracted feature, which encapsulates the semantic essence of the feature being embedded, endowing it with versatility applications.

3. Feature Fusing

- The fusing mechanism is a pivotal component in our models that aims to merge all the feature embedding vectors produced by the previous layer.

Fusing Mechanism module

- Input: The Fusing Mechanism takes input as embedding vectors encoding the following features:
  - Object feature in images (visual features)
  - Enhanced grid features increase the overall contextual understanding.
  - Depth features.
  - Semantic features.
- Output: One final embedding vector retaining the most relevant information from the feature embeddings