RESEARCH AND DEVELOP A DEEP LEARNING MODEL FOR DETECTING AND RECOGNIZING HUMAN ACTION

Nguyễn Đặng Đức Mạnh[1]

[1] University of Information Technology

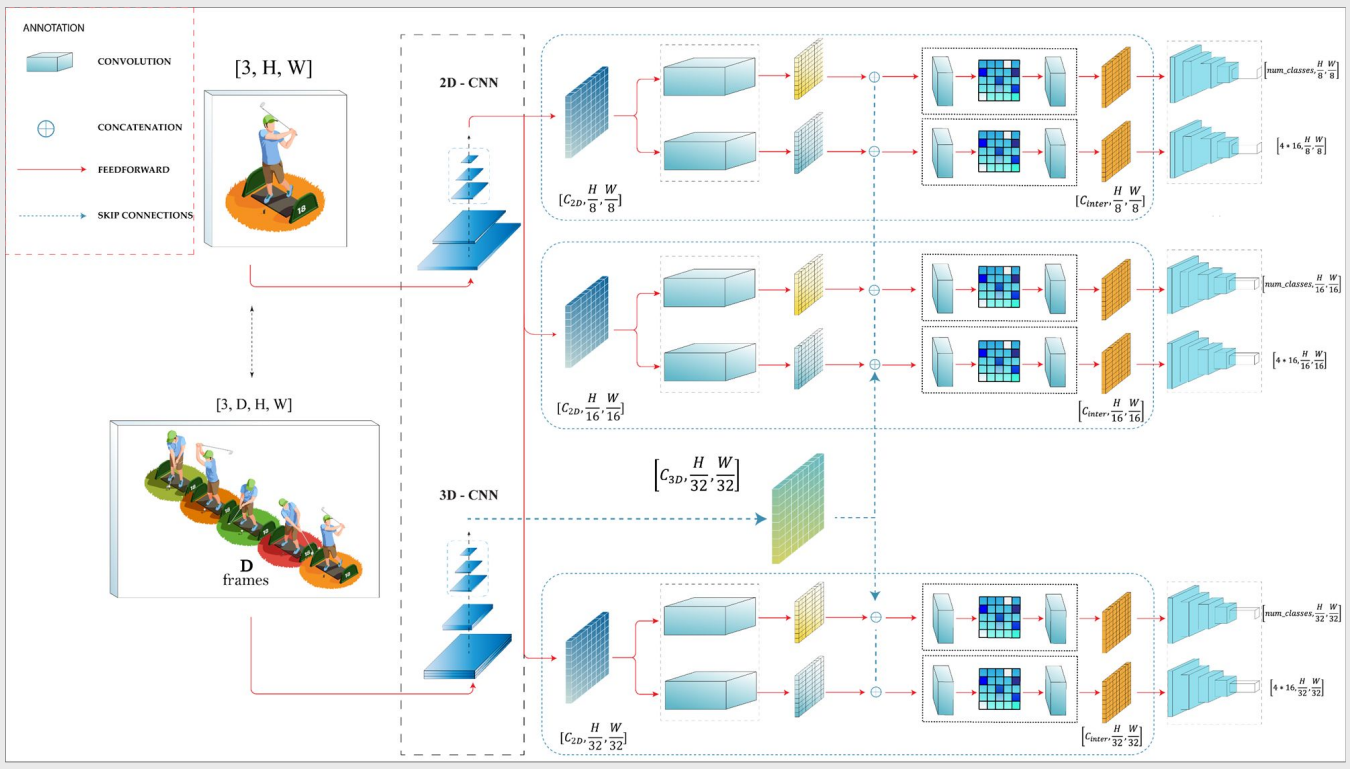
What is HADR?

In the field of computer vision, HADR (Human Action Detection and Recognition) is a task that requires models to detect (bounding box) and recognize (classify) human actions from a given short clip.

Why do we need bounding box?

- **Label**: Useful in video classification tasks, focusing on the general content of the video.
- **Starting + ending point:** Useful in tasks that require querying where a specific action occurs in the video.
- **Bounding box**: Meaningful in more specific query tasks (e.g., person A performing action B), effectively utilizing spatial information in each frame.

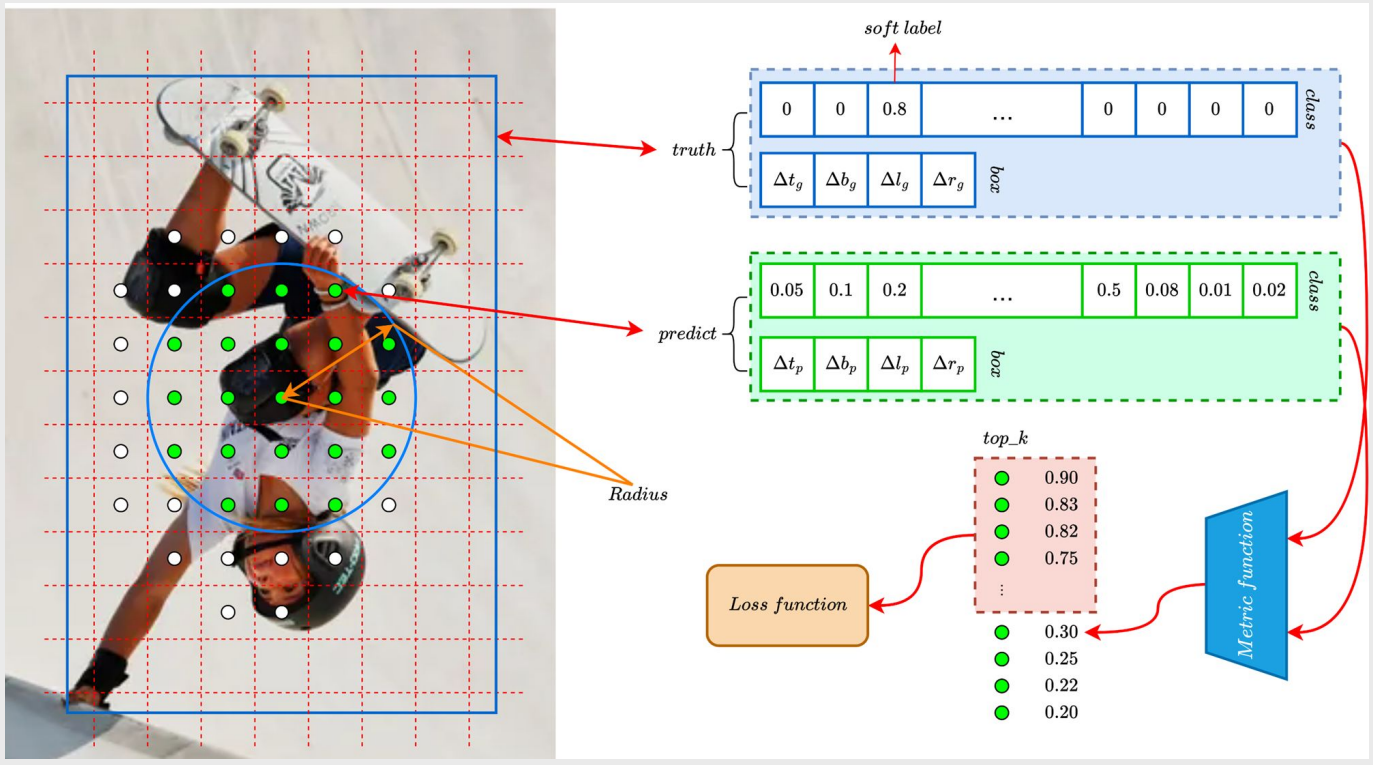
Architecture Overview



- **Backbone 2D**: Used to extract spatial features, we utilize YOLOv8
- **Backbone 3D:** Used to extract motion features over time, we utilize several available 3D backbones including I3D, ResNeXT.
- **Neck:** Enhancing the semantic features of the feature maps by leveraging the Feature Pyramid Network architecture.
- **Head:** Used to predict labels and bounding boxes, the output branch of the model.

Label Assignment

Illustrates in detail the pipeline of the label assignment stage. The candidates deemed eligible (green circles) must lie within the truth bounding box and be within a distance no greater than the radius R from the object's center. Subsequently, the candidates meeting the criteria are evaluated through a metric function, and only the top_k candidates with the highest metric are considered as positive.



NII

Tên tác giả chính– Trường Đại học

TEL : ********* Email : *********