Video Action Recognition With Pytorch In Data-Driven Retail Market

TABLE OF CONTENTS

INTRODUCTION 01

INTRODUCTION

Overview

- The presentation of our Minimum Viable Product (MVP) shows an approach for detecting actions in videos, especially in the retail market's video which is an instrument aimed to assist businesses.
- Action Recognition is widely applied in fields such as surveillance, fitness, sports...etc. Chiefly, through the examination of customer trends and preferences, businesses can formulate targeted marketing strategies aimed at boosting revenue.
- So the question is " How can businesses know if a purchaser will make the decision to buy a specific product among numerous others, rather than just browsing out of curiosity? "

INTRODUCTION

The Problem Statement

The challenge in customer behaviour analysis

In the complex retail market, analyzing buyer behaviour is crucial for businesses to deeply understand their needs and preferences for business development

Time & Effort-intensive

The manual categorization of actions is a process that demands a significant amount of time and effort. This can result in elevated error rates and inefficient use of time during the classification process.

INTRODUCTION

Main Ideas Of The Model

In addition, we have a dataset that includes lists of 2 minute videos with camera overhead and looking down at customers building in a grocery store setting and each video contains numerous instances of 5 actions and our MVP attempts to solve the problem of finding particular actions occurring in a video.

1. Reach To Shelf
2. Retract From Shelf
3. Hand in Shelf
4. Inspect Product
5. Inspect Self

PROPOSED 02

Proposed

Stage 1: Human Action Classification

- Action recognition: Extract feature from each video frame, such as body bose and motion pattern.
- Action classification: A trained action classification model, such as a convolutional neural network (CNN), analyzes the extracted features and classifies each frame into one of the five predefined action categories.

Stage 2: Product Detection & Classification

- Object detection: The system employs object detection technique to identity and localize objects within each video frame.
- Product recognition model: A product recognition model, trained on a dataset of product images, analyzes the detected objects and classifies them into specific product categories.

Proposed

Stage 3: Develop An Integration System Based On A Shop's Dataset

Input: a temporally segmented clip of video is given as input.

Proposed

Stage 3: Develop An Integration System Based On A Shop's Dataset

Output: a modified video containing information about actions and products shown in the video through bounding boxes or action statuses.

METHODOLOGY 03

METHODOLOGY

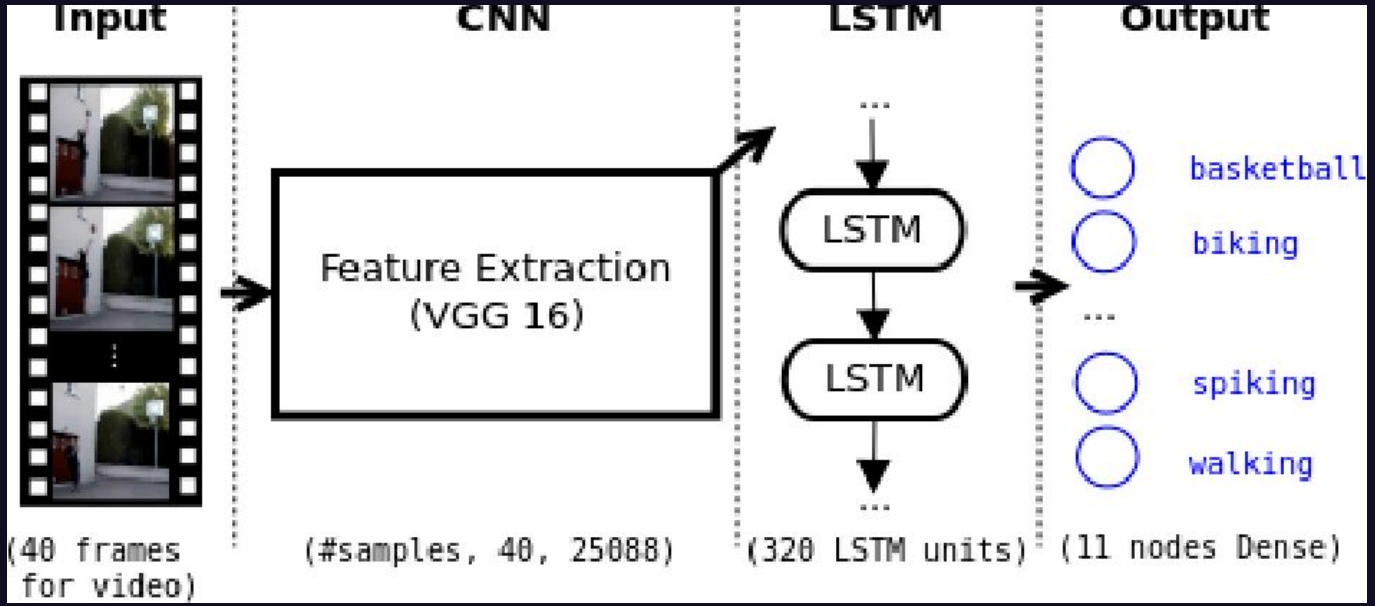

Figure 1: Example of model architecture

METHODOLOGY

Product Detection & Classification

Initially, we will label the coordinate (bounding box) and class name of the product. In this stage we use YOLO, an efficient and accurate classification in object detection.

Illustration for labeling.

METHODOLOGY

Technologies:

- TensorFlow, Pytorch: Deep learning frameworks for building and training neural networks
- OpenCV: A computer vision library for image and video processing
- YOLO: An object detection algorithm
- Front end: HTML/CSS/JS
- Backend: FastAPI

CORE FUNCTIONALITY

04

CORE FUNCTIONALITY

1.) Data Preprocessing

- Extract shot (or frames) from the input videos.
- Segment each video into distinct temporal segments, enduring diverse temporal contexts for training and validation.
- Split each segment into training (80%) and validation (20%)

2.) Feature Extraction

- Employ a pre-trained CNN to extract spatial features from individual video frames
- Utilize transfer learning techniques to leverage a model pre-trained on a large dataset for improve feature extraction.

CORE FUNCTIONALITY

3.) Temporal Information Capture

- Implement a Long Short-Term Memory (LSTM) network to capture temporal dependencies between frames.
- Combine spatial features from the CNN with temporal information from the LSTM to create a robust representation of actions.

4.) Model Training

- Train the combined CNN-LSTM model on the preprocessed dataset

5.) Action Classification Output

- Generate prediction of each frame, classifying action into predefined categories, maybe visualize the action in the video
- Output the result as temporal action sequences for further analysis

CORE FUNCTIONALITY

Product Detection & Classification

Step 1: Labeling of products

- Manually label bounding box coordinate and class names for products in the training dataset

Step 2: Object detection using YOLO

- Implement the YOLO algorithm for efficient and accurate object detection in video frames
- Fine-tune the YOLO model on the labeled product dataset to improve accuracy.

Step 3: Visualization

- Overlay bounding boxes on the video frames to visually represent detected products and actions.
- Output the modified video with annotations indicating recognized actions and detected products.

PERFORMANCE METRICS

05

PERFORMANCE METRICS

Action Recognition Metrics

- Accuracy: the overall accuracy of the model in classifying actions correctly.
- Precision: precision measures the accuracy of the positive predictions. In our context, it would be the percentage of correctly identified actions among all the predicted actions.
- Recall: recall calculates the ability of the model to capture all the actual positive instances. In the retail setting, it presents the percentage of correctly identified actions out of all the true actions that occurred.

PERFORMANCE METRICS

Product Detection Metrics

- Intersection over Union (IOU): IOU measures the overlap between the predicted bounding boxes and the ground truth boxes for product. A higher IoU indicates better localization accuracy
- Precision and Recall for products: similar to action recognition, precision and recall metrics will be used to evaluate the accuracy of product detection. Precision represents the rate of correctly identified products among all the predicted products, while recall represents the ability of the model to capture all the actual positive instances of products.

PERFORMANCE METRICS

Overall System Performance Metrics

- Process Time: measure the time taken by the system to process a given video segment. This is crucial for real-time applications, especially in a retail setting where quick insights are valuable.
- Resource Utilization: evaluate the computational resources (CPU,GPU, memory) consumed during the video processing. Efficient resource usage is essential for scalability and deployment in various retail environments.

PERFORMANCE METRICS

User Satisfaction Metrics:

- User Feedback: collect feedback from users, such as retailers or business analysts, regarding the usefulness and accuracy of the insights provided by the system.
- User Interface Responsiveness: evaluate the responsiveness and user-friendliness of the integrated system’s user interface, ensuring a seamless experience for end-users.

PERFORMANCE METRICS

The Result Of Training, Validation & Test

class	label	total_sample	total_correct_sample	accuracy_per_class
0	No action	10	10	1
1	Reach To Shelf	1	1	1
2	Retract From Shelf	9	8	0.888889
3	Hand In Shelf	18	18	1
4	Inspect Product	26	26	1
5	Inspect Shelf	55	54	0.981818

Accuracy: 0.9832

Figure 2. A training dataset

PERFORMANCE METRICS

The Result Of Training, Validation & Test

class	label	total_sample	total_correct_sample	accuracy_per_class
0	No action	10	1	0.1
1	Reach To Shelf	2	0	0
2	Retract From Shelf	9	6	0.666667
3	Hand In Shelf	21	19	0.904762
4	Inspect Product	26	26	1
5	Inspect Shelf	57	54	0.947368

Accuracy: 0.8480

Figure 3. A Validation dataset

PRODUCT INFOGRAPHICS

06

PRODUCT INFOGRAPHICS

This following figure shows the overall about our MVP. The video output will localize people and recognize action in the whole process of video visualizer (through all frames of video)

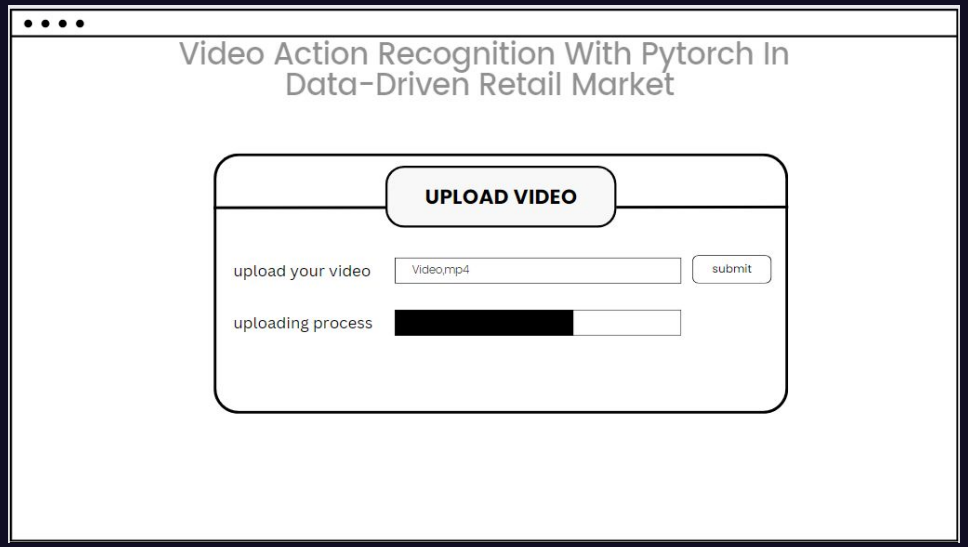

Illustration of User Interface

PRODUCT INFOGRAPHICS

Illustration of User Interface

TIMELINE & ROADMAP

07

TIMELINE & ROADMAP

Date	OBJECTIVE
November 10 - November 12	Received the task and dataset from the organizing committee. Initially processed videos by cutting them into individual frames and fed them into the Resnet, but the results were not promising.
November 13 - November 18	Transitioned to processing videos by shots and switched to the mmaction2 framework.
November 19 - November 26	Aggregate various methods for video data processing and learn how to work with the mmaction2 framework for action recognition
November 27 - December 1	Brainstorm ideas to finalize the approach for data processing and learn how to deploy AI applications on web platforms.
December 2 - December 3	Participate in a workshop and training session organized by a mentor.
December 4 - December 15	Teamwork with the mentor to prepare for Vietnam Datathon Day.
December 16 - December 17	Participate in Vietnam Datathon Day.

OPPORTUNITIES & LIMITATIONS

08

OPPORTUNITIES & LIMITATIONS

Weakness

Limited Context Analysis

While computer vision excels at identifying actions and objects in individual frames, it struggles to capture the nuances of customer behavior that span multiple frames. For instance, it cannot directly infer the relationship prolonged product inspection and a higher purchase likelihood.

Inflexibility In Adapting To new Trends

The system's reliance on computer vision alone limits its ability to adapt to emerging customer behaviors, product trends, and technological advancements. Ongoing maintenance and updates are crucial to ensure its effectiveness in a dynamic retail environment.

Challenges In Integrating With Existing System

Integrating the system with existing retail infrastructure and business processes may require significant effort and customization. This could hinder the seamless flow of data and hinder the system's ability to provide actionable insights to businesses.

OPPORTUNITIES & LIMITATIONS

Future Potential

OPTIMIZATION

Expanding our model could involve incorporating features like video summarization to detect customer presence in-store, optimizing memory usage, and improving operational efficiency by eliminating periods with no customers, such as during adverse weather conditions or low foot traffic.

ANTI-THEFT

Addressing the issue of merchandise loss due to theft in retail stores, we aim to develop a product that analyzes customer behavior, counts selected items, and cross-references this data at the checkout counter to identify potential theft, reducing the amount customers pay during their shopping experience.

CONCLUSION 09

CONCLUSION

Summary

- Our model not only discerns actions within the dataset for insightful analysis of customer behavior but also offers a user-friendly interface for practical use in sales.
- Accuracy, precision, and recall for classifying actions correctly.
- Measures process time, resource utilization, and user satisfaction through feedback and interface responsiveness.
- Furthermore, future development efforts will concentrate on enhancing the model's capability to recognize a broader range of customer actions, providing additional advantages for businesses.

THANKS!

REFERENCES

[1] https://github.com/open-mmlab/mmaction2

[2] https://github.com/IBM/action-recognition-pytorch

[3] https://github.com/open-mmlab/mmengine

[4] https://github.com/soCzech/TransNetV2