Method description

Presented by: Amit Hayun, Bar Loupo
Lecturer: Dr. Marina Litvak

1.Task definition

Our goals are to recognize abuse events in real-time through an IP camera, using deep neural network architecture.
The input to the model is a 5 second video clip represented as an np array of size (224,224,3).
The output of the model is a prediction for the clip if it contains abuse or not.

2.Theoretical background of our approach

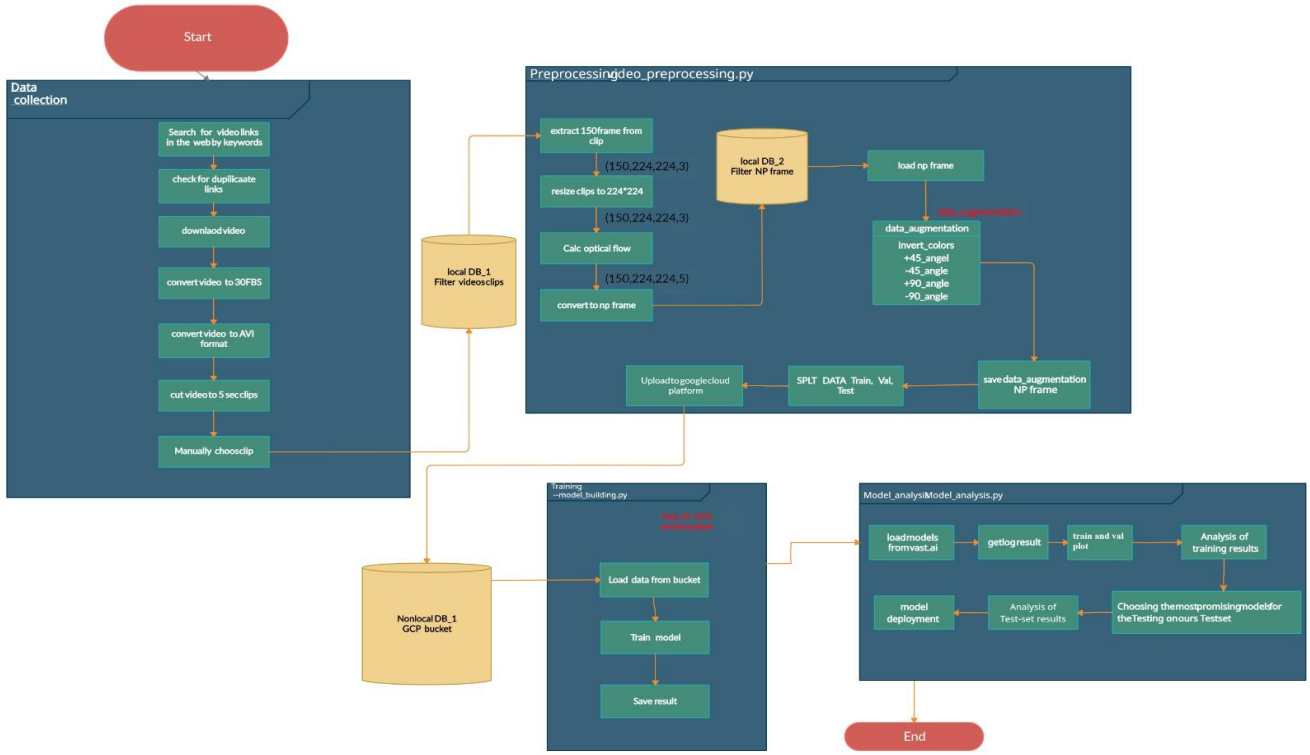
Today most computer vision recognition tasks use obvious choice Neural Network architecture such as 3DCNN, CNN, ConvLSTM, SVM.
Our demand for real-time detection forces us to consider the size of models, we want the smallest architecture that allows us to make predictions on the real-time system.
Our approach was to use supervised learning with 3 NN models that uses advanced algorithms to detect and track objects.
The NN models that we used are:

- YOLO_v3 – a fully convolutional network for detecting objects in frames.
- DeepSORT – a NN model for tracking objects in video
- Gate_flow – a NN model for violence detection.

Stages of our approach:

- Stage 1: is to input a video in batches of frames to the YOLO_v3 model that draws a bounding box around the objects detected in them.
- Stage 2: is to input the frames outputted from YOLO_v3, with their bounding boxes, to the DeepSORT model which tracks the detected objects in the frame and assigns each on of them a unique ID.
- Stage 3: is checking if two people were detected in a frame, if not we return to stage one – YOLO_v3, if yes, we measure the distance between them and if it is less then or equal to a meter we continue to the next stage, otherwise we return to stage one – YOLO_v3.
- Stage 4: two people were detected in a frame and the distance between them was less then or equal to a meter, we preform preprocessing operations on the frame and input it to the gate_flow model which outputs a prediction of abuse/not abuse.
- Stage 5: if the gate_flow model predicted abuse, we save the video containing the abuse and send it to the relevant parties. if the gate_flow model predicted not abuse we release the frames and return to stage 1.

3.Graphical pipeline



4.Experiments

- Data collection:
  We searched the internet for datasets for violence and abuse detection, adult care, and normal behavior(we searched for routine activities that take place at home on a regular basis). In addition, we searched YouTube for videos of abuse in several languages. We combined all the videos we found into one main dataset which we used to train validate and test our model on. The dataset consists of:
  - **Abuse of the elderly** - More than 200 videos. After manual sorting, we extract 160 video clips of elder abuse
  - **child abuse** - More than 150 videos. After manual sorting, we extract 70 video clips of child abuse, most of the video clips are from CCTV.
  - **Street Fight** - This data set is from **NTU CCTV-Fights Dataset** and contains 1000 videos. We extract 250 Fight clips that are taken from a mobile camera.
  - **Adult care** - More than 200 videos. After manual sorting, we extract 175 video clips that contain Changing a diaper, dressing, feeding, nursing, lifting, etc.
  - **Normal behavior_1** – More than 150 videos. After manual sorting, we extract 100 video clips of Normal behavior such as Lifting a sofa, moving furniture, Walking at home, etc.
  - **Normal behavior_2** - This data set is from the Real-Life **Violence Dataset. We extract** 150 video clips of normal behavior such as People eat, talk, lift things, human contact.

  Overall, the dataset contains 905 videos.

- Data preprocessing:
  The data preprocessing was in the following steps:
  1. convert the video to AVI format with FBS=30sec
  2. cut the video into 5 seconds clips
  3. manually extracting from each video 3 or 4 clips
  4. use 6 data augmentation techniques to create more videos and enlarge the train portion of the dataset
  5. split the data to Train, Validation, Test as shown in tables 1 and 2

Table 1: Dataset partition

Source	Train	Validation	Test
Abuse of the elderly	120	20	25
Child abuse	50	10	0
Street Fight	156	40	0
Adult care	127	20	25
Normal behavior_1	100	0	0
Normal behavior_2	100	50	0

Table 2: Abuse / Not Abuse distribution

	Abuse	Not Abuse	Total
Train	326	326	652
Validation	70	70	140
Test	25	25	50

- Hyperparameters:

  Experiment 1 – SGD:

learning rate α	**0.01**
learning decay	**1e-6**
momentum β	**0.9**
batch size	**6**
Number of workers[FOR training]	**6**
Number of epoch	**30**
GPU	1x Tesla V100

Experiment 2 – ADAM:

learning rate α	**0.01**
epsilon ε	**1e-07**
beta $\beta 1$	**0.9**
beta $\beta 2$	**0.999**
batch size	**6**
Number of workers	**6**
Number of epoch	**30**
GPU	1x Tesla Titan RTX

- Hypotheses/Ablation study:
  The Adam optimization algorithm has shown better performance on our test set than the previous SGD optimization algorithm.
  We can see an improvement of 5% in F1-Score and an 8% improvement in recall.

Best Epoch	Train Accuracy	Validation Accuracy	Test Accuracy	Recall	Precision	F1-Score
SGD – Epoch 19	0.89	0.829	0.8	0.68	0.89	0.77
ADAM – Epoch 24	0.986	0.921	0.84	0.76	0.905	0.826

We can see that the model gave the best accuracy of 84% and overall better performance on all measurements with the ADAM optimizer.

- Results and their interpretation:

    - The model classifies video clips as **violent** even when one person makes a sudden movement and there is no proximity between the two people in the video, which makes a lot of sense, since the model is looking for changes in the optical spacing between 2 neighbouring frames. Therefore, any fast action, regardless of the distance between two people, will cause a significant change in the optical current. This is a very big problem for our system, because most of the nurses/caregivers in a nursing home perform additional activities daily, for example folding laundry, changing bedding, etc.

    - The model has difficulty analyzing frames that contain a TV/computer screen. The sharp movement of the frames on the TV causes a high result in the optical current thus misleading the model and causing it to classify the event as abuse. Moreover, because in most nursing homes / private homes the camera is usually aimed at the living room and/or bedrooms that almost always have a TV, it can be misleading for the model.

5.Future work

This paper presents both a novel dataset and a method for abuse detection in videos.
In the future, to train a more robust and reliable model, the size of the dataset needs to be further expanded.
Also, in future further methods of data augmentation could be explored in order to expand the current dataset.